Universidad Carlos III de Madrid
Escuela Politécnica Superior

Ingeniería de Telecomunicación

Proyecto Fin de Carrera

Sistema de Question Answering basado en Wikipedia

Autor: Jesús Fernández Benito
Tutor: Julio Villena Román

14 de julio de 2006

A mi madre

Agradecimientos

En el momento en el que atisbé que esta etapa de mi vida, agridulce e irracional como sólo sabe serlo la universitaria, estaba tocando a su fin, aparecieron ante mí dos sentimientos bien distintos. Por un lado, alegría ante la perspectiva de salir de un período en el que últimamente me sentía atado, descolocado y sin rumbo. Por otro, me invadió, sin que me diera cuenta, una sensación que todavía no podría definir, pero que, por darle un nombre, voy a llamar *miedo*.

"Y en ese mismo instante en que el mundo que lo rodeaba pareció desvanecerse y él se quedó solo como una estrella en el firmamento, en aquel momento de frialdad y de desánimo, se irguió un Siddharta más sólido y fuerte, más posesionado que nunca de su propio Yo. Se dio cuenta de que aquello había sido el último estremecimiento del despertar, el espasmo final del parto. Y al punto reanudó su marcha, con paso rápido e impaciente…; mas no a su casa […]; ya no hacia atrás."

Siddarta. **Herman Hesse**

Gracias a este "renacimiento", y apoyándome en dos grandes filósofos de la vida, *Beppo Barrendero* y *Fuckowski*, seguí hacia delante. Recordé entonces que hacía mucho tiempo que no revisaba los libros de contabilidad, así que me puse a hacer balance: lo bueno y lo malo, ilusiones ganadas, perdidas y olvidadas, de las amistades que tenía, en las que creí y las que quedan… y comprendí que, en realidad, salía ganando. Sin embargo, existe un apartado en estos libros que está escrito, en su mayor parte, con grandes y numerosos números rojos, es la sección de Agradecimientos.

En este apartado, el mayor descubierto le tengo con mi familia. Ese universo (que otros llaman "casa") se compone de número indefinido de galerías, no necesariamente infinito ni de forma hexagonal, en el que habita un conjunto variopinto de personas definidas a la perfección por un funcionario suizo.

"Al llegar a la frontera suiza, un funcionario […] inspeccionó nuestros equipajes. Se los devolvió a Mamá junto con una hojita de papel, […] en la columna titulada ‹‹Descripción de los Pasajeros››, aparecía escrito en pulcras letras mayúsculas: Un Circo Ambulante y su Compañía."

Mi familia y otros animales. **Gerald Durrell**

Es a ese grupo de personas (con las que no se habla, se discute) al que nunca seré capaz de agradecer el gran esfuerzo diario que solo ellos son capaces de realizar, por ejemplo, en la difícil tarea de soportarme.

No obstante, tengo descubiertos de gratitud en un número elevadísimo de cuentas. La vida me ha regalado conocer a personas muy interesantes, otras menos, que me han aportado, aún sin notarlo, algo de sí mismas que me ha convertido en lo que soy. Por ello, sería injusto agradecer aquí sólo a las personas que, en este momento, considero más importantes (si ellas no lo saben, algo estoy haciendo mal). No sois pocos los que me habéis regalado infinidad de cosas que tendría que agradecer: unos apuntes, tu compañía, una cerveza, hacer soportables los meses de "prisión" en la biblioteca, un viaje, un abrazo, una conversación absurda, una ilusión, la bajada de una pista, tu cariño, mis pelos de punta en un concierto, un Jameson (solo, 2 hielos), tu preocupación, una sonrisa, desayunar sin acostarme, una ola, tu apoyo, una película, un beso, alguna visita de "el del peto azul", el hacer posible este proyecto, ese momento sólo nuestro… Tantas cosas... sin ellas, habría abandonado, o habría terminado en la mitad de tiempo.

También existen otras personas, aquellas que, al contrario que las anteriores, están deseando que caigas, ponen trabas, nos desesperan, amargan, que intentan quitarnos la ilusión… a ellas también les debo, ya que ahora soy más fuerte.

"Mis enemigos quisieran verme bien muerto, a ellos les dedico cada uno de mis éxitos"

No me convences. **Jacky Trap**

Llegados a este punto, aunque no creo que sea éste el momento más indicado para agradecer todo lo que debo, y me parezca tarde o a destiempo, John Ruskin dijo: *"Lo que creamos o lo que pensemos, al final no tiene mayor importancia. Lo único que realmente importa es lo que hacemos."* y voy a hacerle caso, no voy a dejar que se me pase otra oportunidad, a todos vosotros, pero principalmente para los buenos, los bellos, los sencillos, aquellos que os distinguís por unos buenos ojillos…

Gracias ☺

ysus.
Madrid, 2 de julio de 2006

Resumen

"Si he aprendido algo, es que todo puede saberse. Sólo tienes que encontrarlo."

Sueña conmigo (The Sandman #3). **Neil Gaiman**

Este proyecto se centra en la investigación de las técnicas y estrategias que se emplean actualmente en los sistemas de respuesta automática, más conocidos por su denominación inglesa Question Answering (QA), disciplina que forma parte de la Ingeniería Lingüística (rama de la Inteligencia Artificial encargada del estudio y procesado del lenguaje natural) en la que se diseñan sistemas capaces de interpretar preguntas que realizan los usuarios, para buscar los documentos relacionados, extraer la información solicitada y devolver una respuesta completa.

Utilizando esta investigación como base, se plantea el diseño y arquitectura un sistema de QA genérico, para su posterior implementación, teniendo en cuenta especialmente las características de la lengua española que requieren un tratamiento diferenciado de otros idiomas, al que se ha exigido un nivel de aciertos comparativamente similar a los sistemas actuales de QA y una interfaz que permita su utilización a personas sin conocimientos técnicos. La implementación efectiva del sistema se divide en dos fases: adquisición del conocimiento e interacción con el usuario.

En la primera fase, de adquisición del conocimiento, el sistema descarga, procesa e interpreta los artículos de Wikipedia, la enciclopedia libre. Mediante este procedimiento, el sistema incorpora estos artículos a su "cultura" y se prepara para poder contestar las preguntas que se le formulen. En la segunda fase, de interacción con el usuario, cada vez que recibe una pregunta, la procesa, la analiza y busca información relacionada en su base de datos. Después extrae las posibles soluciones, las examina, las clasifica por relevancia y muestra las mejor valoradas al usuario.

En las pruebas de evaluación, el sistema básico logra dar una respuesta apropiada al 22% de las preguntas y, tras la adición de dos bloques de expansión (utilización de sinónimos en la búsqueda y categorización de las respuestas en función del tipo de pregunta), este porcentaje sube hasta el 27'5%.

Abstract

"One thing I've learned: You can know anything. It's all there. You just have to find it."

Dream a little dream of me (The Sandman #3). **Neil Gaiman**

This project focuses on current existing methods for building Question Answering (QA) systems, an interdisciplinary field related to the Information Retrieval and Natural Language Processing areas whose objective is to develop systems which are able to automatically provide correct answers to questions posed by users.

The objective of the project is to propose, after an exhaustive preliminary research, the architectural design of a basic QA system and then develop an actual system with three main design guidelines: the system must take into account the special characteristics of the Spanish language, must achieve similar results (in terms of correctly answered questions) to other existing systems and, finally, must provide an "easy to use" web interface.

The system runs in two phases: the learning process and the answering process. First, Wikipedia must be parsed to extract its meaningful text fragments, which are then splitted into sentences to finally "learn" all this knowledge by means of an Information Retrieval engine. In the second phase, the same processing steps are followed to try to provide the users with a valid answer (or more than one) for their questions: first of all, the question is parsed and POS-tagged; then, the search engine is used to find the Wikipedia sentences which are most related to the question and are supposed to contain the answer; and, finally, after extracting, ranking and sorting those sentences, the best choices are shown to the users in a friendly web interface.

The evaluation shows that the basic system achieves a 22% of correct answers and, after the addition of two specialized modules (expansion with synonyms and question-type classifier), this percentage increases up to 27.5%, which is in fact a very good rate compared to existing systems.

Índice General

Índice de Figuras

Índice de Tablas

Capítulo 1.

Introducción

"- Pero, ¿por dónde empiezo, Simeón?
- Empiezas por una elección."

La paradoja. **James Hunter**

1.1. Motivación del proyecto.

La Inteligencia Artificial, disciplina iniciada por Alan Touring[1] en su artículo "*Computing Machinery and Intelligence*" (publicado en 1950), aunque no empezó con muy buen pie y estuvo casi abandonada durante años, hoy en día se ha convertido en algo cotidiano, y se encuentra inmersa en la vida diaria, oculta tras sistemas de reconocimiento, motores de búsqueda (e.g. *Google*[2]) o videojuegos. Una de las áreas con mayor desarrollo en este campo es la Ingeniería Lingüística, encargada del estudio y tratamiento del lenguaje natural (el humano); dentro de esta disciplina se puede catalogar a un amplio conjunto de sistemas, correctores ortográficos, analizadores lingüísticos o traductores automáticos.

La Recuperación de Información (RI, o IR del inglés *Information Retrieval*) es una de las ramas de la Ingeniería Lingüística, en ella se investigan y desarrollan los motores de búsqueda, sistemas capaces de devolver una lista de documentos (páginas Web, imágenes, texto…) en relación con unas características específicas definidas por el usuario por medio de una consulta.

[1] Matemático británico, científico, criptógrafo y filósofo. Uno de los padres de la informática moderna, diseñador del primer computador electrónico digital y funcional, COLOSSUS, utilizado descubrir códigos secretos nazis (como los de la famosa *Enigma*) durante la Segunda Guerra Mundial.
[2] http://www.google.com/

El Procesado del Lenguaje Natural (PLN, o NLP del inglés *Natural Language Processing*), otra de las ramas de la Ingeniería Lingüística, estudia la manera de que una máquina sea capaz de "entender" el lenguaje humano; abarca una amplia gama de categorías, desde la traducción a la generación del habla o el auto resumen.

Entre la RI y el PLN, encontramos los sistemas de respuesta automática, más conocidos por su denominación inglesa Question Answering (QA), sistemas capaces de interpretar una pregunta formulada por el usuario, buscar en su interior documentos relacionados y extraer, de sus contenidos, una respuesta concreta, clara y específica para la pregunta introducida.

Dentro del marco de desarrollo actual, en el que la informática trata de simplificar y automatizar procesos al ser humano, ofreciendo a la vez mayor facilidad de uso, los sistemas de QA aparecen con el objetivo de resolver cualquier tipo de pregunta que le pueda surgir a un usuario, acercándose a él de una manera fácil y eficaz. Esta circunstancia, junto con la cobertura ofrecida por los sistemas de código abierto, fueron los detonantes principales de la puesta en marcha de este proyecto.

1.2. Objetivos.

El objetivo principal de este proyecto es la realización de una investigación exhaustiva de las técnicas y procedimientos más utilizados en la construcción de sistemas de QA, como base para poder plantear el diseño y la posterior implementación completa de uno de estos sistemas, cumpliendo una serie de requisitos:

- Deberá utilizar como lengua de trabajo el español (castellano).
- El nivel de aciertos deberá ser comparativamente similar a los sistemas actuales de QA.
- Se exigirá un diseño modular, para facilitar las labores de investigación, permitiendo de una manera sencilla la adición o sustracción de bloques
- Su interfaz de uso tendrá que ser muy clara e interactiva, para ser usada por cualquier persona sin conocimientos técnicos, desde la web.

1.3. Contenido de la memoria.

En los siguientes capítulos de este documento se describen todos los aspectos de interés que tienen relación con el proyecto, desde el estado actual de la tecnología implicada hasta los resultados y conclusiones obtenidas tras la implementación, pasando por los aspectos más relevantes del desarrollo del proyecto.

- **Capítulo 1: Introducción,** el presente capítulo.
- **Capítulo 2: Sistemas de QA**, donde se describe el estado actual de las tecnologías implicadas en el desarrollo del proyecto.
- **Capítulo 3: Descripción del Sistema**, capítulo en el que se describe la arquitectura y funcionamiento del sistema implementado.
- **Capítulo 4: Adquisición del Conocimiento**, encargado de explicar el tratamiento que se ha dado a la información hasta que llega a ser parte del sistema.
- **Capítulo 5: Contestando Preguntas**, donde se describe el esquema de funcionamiento del sistema, desde la introducción de la pregunta hasta la devolución de las respuestas.
- **Capítulo 6: Pruebas**, lugar en el que se describirán los resultados obtenidos tras la implementación.
- **Capítulo 7: Conclusiones y Trabajos Futuros**, donde se exponen las conclusiones obtenidas y se recogen las posibles líneas futuras de desarrollo.
- **Apéndice A: Preguntas QA@CLEF 2006,** preguntas utilizadas en las pruebas.

Capítulo 2.

Sistemas de QA

"No conozco las suficientes palabras –se dijo-. Y hay cosas que uno no puede pensar a menos que conozca las palabras adecuadas."

Camioneros. **Terry Pratchett**

2.1. Introducción.

El objetivo de este capítulo es el de dar una visión global de conceptos que forman parte de las bases de este proyecto, describiéndolos y mostrando su estado de desarrollo actual. Se empezará con una explicación general de los Sistemas de Question Answering y, partiendo de ella, se hará un examen exhaustivo de sus partes, mostrando ejemplos y referencias de sistemas reales para facilitar la comprensión.

2.2. Question Answering.

Los sistemas de QA son aplicaciones informáticas que son capaces de contestar correctamente a preguntas que se han formulado sin tener en cuenta que van dirigidas a máquinas. Es decir, sistemas que pueden responder, directamente, a preguntas como "*¿Quién es el Ministro de Industria?*" o "*¿Dónde nació Fernando Alonso?*" de una manera clara y concisa, como "*José Montilla Aguilera*" o "*en Oviedo*".

2.2.1. Descripción de un sistema de QA.

El esquema general de estos sistemas suele dividirse en tres grandes bloques, unidos de manera secuencial:

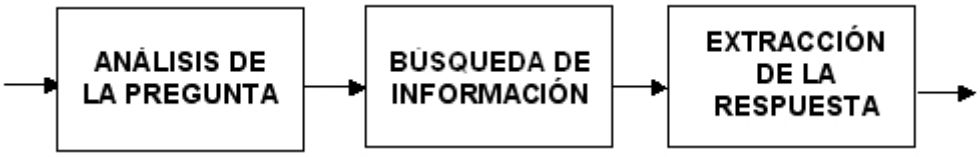

Figura 1. *Esquema de bloques de un sistema de QA.*

Las características principales de estos bloques son:

1. **Análisis de la Pregunta:** Este módulo es el encargado, mediante la utilización de herramientas de PLN, de transformar la pregunta en una *consulta* que entienda el siguiente bloque. Además puede sacar información de la pregunta, como puede ser el "tipo de pregunta", que sea útil a la hora de elegir una respuesta.

2. **Búsqueda de información:** Habitualmente, un sistema RI de recuperación de texto (documentos, párrafos…) que procesa la *consulta* que le pasa el bloque anterior y devuelve una colección de documentos o fragmentos de ellos (páginas, párrafos, frases...) relevantes, relacionada con la cadena introducida.

3. **Extracción de la Respuesta:** En este bloque se trocea la información obtenida en el apartado anterior, se filtra lo que puede ser una respuesta y se evalúa, por diferentes métodos, la probabilidad de que sea la respuesta correcta; devolviendo la mejor valorada.

2.2.2. Características principales.

Las características deseables para estos sistemas fueron definidas en 2002 por un grupo de investigadores en una guía (véase [7]) de cómo deberían evolucionar las técnicas de QA y de cómo lo ha ido haciendo hasta ahora. Según ese documento, estas características son:

- **Independiente del tiempo (***Timeliness***):** Una respuesta tiene que poder ser contestada aunque se refiera a acontecimientos muy recientes. Además, el sistema debe proporcionar respuestas en tiempo real, interactividad.

- **Exactitud (***Accuracy***):** El sistema debe ser todo lo preciso y exacto que se pueda, ya que si una respuesta es incorrecta puede causar muchos problemas. Una respuesta incorrecta es mucho peor que una pregunta no contestada.

- **Utilidad o "usabilidad" (***Usability***):** Debe ser capaz de dar una respuesta en el formato deseado por el usuario, independientemente del tipo de archivos que tenga que manejar.

- **Completitud (***Completeness***):** Es deseable que la respuesta sea completa. Por lo tanto, si ésta se haya distribuida por distintos documentos, es necesaria una fusión de todas las partes antes de presentársela al usuario.

- **Relevancia (***Relevance***):** Debe dar respuestas relevantes dentro de un contexto. Si no es posible la deducción, se debe preguntar al usuario. Se necesitan sistemas interactivos.

2.2.3. Problemas principales.

Los problemas principales a los que se enfrentan estos tipos de sistemas van a producirse a la hora de la interpretación del lenguaje, donde se generan dificultades de muy diversa naturaleza.

Un par de problemas típicos pueden ser: la dificultad en la elección de un contexto adecuado, cuando éste no está explícitamente en la pregunta (por ejemplo, en el caso de preguntar "*¿Quién es el presidente del Gobierno?*" falta por definir el contexto del país) y los problemas para identificar un mismo tipo de pregunta, independientemente de cómo se haya formulado (ejemplo: "*¿Cuánto cuesta…?*", "*¿Qué precio tiene…?*").

Además, el porcentaje de acierto que va a tener un sistema de QA no va a depender tanto de la dificultad de la pregunta sino de cómo encaje ésta con la respuesta encontrada. Esto puede verse en el siguiente ejemplo:

P: *¿Quién descubrió América?*

T1: *12 de octubre de 1492: el día en que Cristóbal Colón descubrió América*[3]*.*

[3] Fuente "Sí España": http://www.sispain.org/spanish/history/discover.html

T2: *La frase descubrimiento de América se usa para referirse a la primera llegada de españoles a América con consecuencias históricas, la de Cristóbal Colón a una isla del mar Caribe*[4].

Si la similitud entre la pregunta y los textos en los documentos disponibles es tan clara como en el caso de **P** y **T1**, un sistema de QA no encontrará mucha dificultad en encontrar la respuesta correcta. Sin embargo, cuando la similitud es tan vaga como entre **P** y **T2,** los problemas para encontrar la respuesta aumentan.

2.2.4. Desarrollo mundial.

Como estímulo para potenciar estas tecnologías, nació en 1992 TREC (*Text REtrieval Conference*)[5] en Estados Unidos, patrocinado por el NIST (*National Institute of Standards and Tecnology*) y por el Departamento de Defensa. En un esfuerzo por crear una comunidad de investigadores y desarrolladores que estudien y desarrollen sistemas RI.

En Europa ha surgido otro centro de reunión, el CLEF (*Cross-Language Evaluation Forum*)[6], promovido por entidades públicas y privadas de toda Europa y del resto del mundo. Con objetivos similares a los de TREC, pero con el añadido de dar mucha más importancia a la interconexión y diversidad de las lenguas. Busca, además, impulsar los sistemas europeos y así garantizar su competitividad en el mercado global.

Estas concentraciones son un lugar de reunión para los investigadores en el campo de la RI. Aquí, los participantes someten a sus sistemas a una serie de pruebas diseñadas para la ocasión, luego evalúan los resultados, y por último todos comparten sus experiencias. De esta manera se crea un lugar en el que es posible intercambiar ideas y comprobar sus resultados, consiguiendo así cumplir los objetivos que se marcaron en un principio:

- Animar a la participación en la investigación de este campo.
- Crear unas comunicaciones entre las principales partes implicadas en el proceso (investigadores, gobierno e industria) más continuas y directas.
- Potenciar la transferencia de tecnología, aumentar la velocidad de conversión de los productos de laboratorio en productos de mercado.

[4] Fuente "Wikipedia": http://es.wikipedia.org/wiki/Descubrimiento_de_América
[5] http://trec.nist.gov/
[6] http://clef.isti.cnr.it/

- Permitir a todos una mayor disponibilidad de técnicas de evaluación para mejorar el desarrollo.

Gracias a estas conferencias, cada año los sistemas mejoran (aumenta su precisión, su cobertura, eficiencia...), el número de investigadores aumenta, así como el número de países implicados. Se van añadiendo categorías, y se desarrollan las existentes.

2.2.5. QA bilingüe.

Hasta ahora, los sistemas de QA que se han descrito en este capítulo han sido considerados monolingües, es decir, que el idioma de las preguntas y el de la colección de documentos es el mismo. Dado que el volumen de los recursos de información disponible en Internet en algunas lenguas es muy escaso (ver **Figura 2**), existen sistemas, como el desarrollado en [11], que son capaces de buscar información en lenguas distintas a la de formulación de la pregunta.

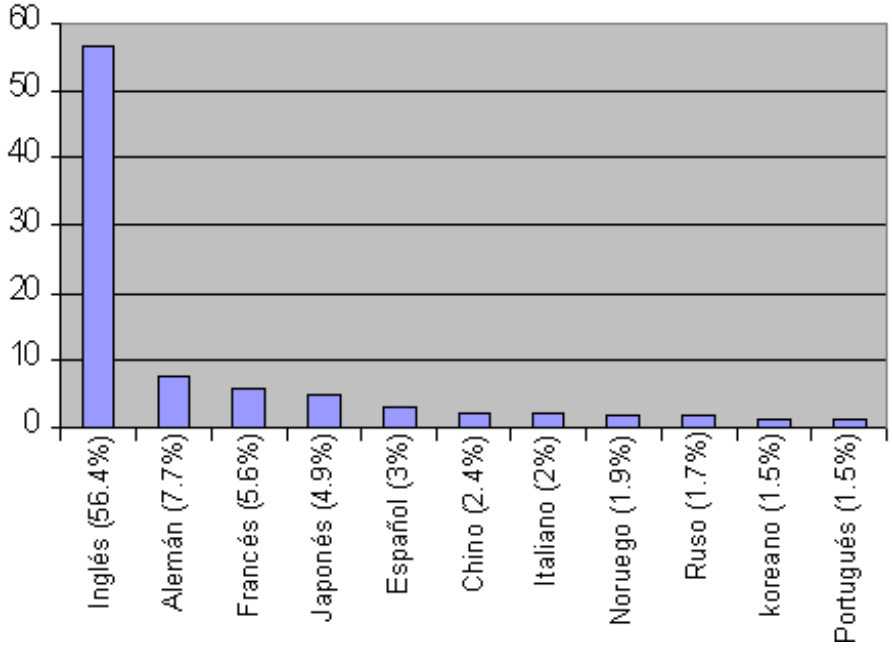

Figura 2. ***Porcentaje de páginas web por idioma.***[7]

Para desarrollar estos sistemas es necesario incluir un módulo traductor. Las dos soluciones que se presentan en un principio son o bien la traducción de los textos que se utilizan como base del conocimiento, o la traducción de la pregunta y de la respuesta. De esta manera se conservarían sistemas monolingües con módulos de traducción.

[7] Fuente "Netz-Tipp-Studie": http://www.netz-tipp.de/sprachen.html

De estas dos, la opción más razonable es la opción de colocar un bloque traductor a la entrada y otro a la salida. Pero, por el momento, los sistemas traductores tienen problemas, no son exactos, por lo que la parte de análisis de la pregunta queda afectada.

Para intentar reducir estos problemas, surge una solución que se basa en analizar la pregunta en el idioma origen y traducir después las palabras de la consulta, con lo que evitarían algunos problemas de interpretación.

Pero al final, hay errores que permanecen, y que no tienen una solución sencilla. Algunos de los más típicos se producen:

- En la traducción de palabras clave: provocados por la falta de contexto a la hora de la traducción.
- En la traducción de nombres: ya que no existe un criterio con respecto a la traducción de los nombres. Los de personas y compañías no suelen traducirse (por ejemplo, "*Michael Schumacher*", "*Miguel Indurain*"), pero sí los de ciudades o países ("*Londres*"/"*London*", "*España*"/"*Spain*").
- En la traducción de siglas: debido a que éstas suelen depender siempre del lenguaje de origen, para poder traducirlas es preciso que primero se conozca su significado completo.

2.3. Análisis de la pregunta.

En este apartado se tratará de explicar la importancia del bloque de Análisis de la pregunta, comentando sus diferentes etapas y posibilidades.

2.3.1. Primer análisis y etiquetado.

El primer paso es realizar un análisis a "bajo nivel" de la pregunta. En este primer análisis se realiza la *segmentación* (en inglés, *tokenization*), encargada de identificar las palabras simples y las compuestas, para después clasificarlas y etiquetarlas según su categoría morfológica (*Part-Of-Speech Tagging*), lo que facilitará su posterior tratamiento.

Como ejemplo de este tipo de sistema, puede verse el sistema [9], llamado **MBT** (*Memory Based Tagging*) y que dispone de una "demo" accesible vía web[8].

Una vez realizado éste análisis, existen sistemas que deciden pasar al siguiente módulo. A partir de este punto, se pueden elegir las palabras que se buscarán, teniendo en cuenta su categoría morfológica (por ejemplo, un nombre tendrá más importancia que una preposición). Pero se ha comprobado que se obtienen mejores resultados si se realiza un análisis más profundo de la pregunta y se utiliza la información obtenida para mejorar los parámetros de búsqueda y los criterios de extracción de respuestas.

2.3.2. Análisis del tipo de pregunta.

En un nivel más profundo de análisis, lo primero que se debe identificar es el *tipo de pregunta*, ya que si se conoce el tipo de pregunta se puede acotar el campo de búsqueda, centrándolo en la categoría semántica de la respuesta. Además, se puede utilizar para crear distintos métodos de actuación en función de la pregunta a la que se enfrente el sistema.

Cada implementación identifica unos tipos de pregunta, dependiendo de los sistemas de análisis de que se dispongan y de los criterios de cada desarrollador. Cuantos más tipos se diferencien, más específicos serán estos, y mayor va a ser la reducción del campo de búsqueda de la respuesta. Esto puede verse en la **Tabla 1**, donde se compara un sistema con una categoría general con otro que tiene categorías más específicas.

Tabla 1. *Ejemplo de tipos de pregunta.*

general	**tipo de pregunta**	**entidad de la respuesta**
	CUÁNTO/A/OS/AS	*<cantidad>*
específico	**tipo de pregunta**	**entidad de la respuesta**
	CUÁNTO-CUESTA *CUÁNTOS-KILÓMETROS* *CUÁNTO-PESA* *…*	*<cantidad monetaria>* *<cantidad-distancia>* *<cantidad-masa>* *…*

[8] http://ilk.uvt.nl/~zavrel/tagtest.html

El aumento del número de categorías identificadas mejora la funcionalidad de los sistemas. Sin embargo, la cantidad de categorías que se pueden llegar a catalogar es muy grande. Por ello, se han desarrollado otras vías para enfrentarse al problema de la entidad de la respuesta.

Una de ellas es la utilización del *foco* (en inglés, *focus*). El foco es una palabra o conjunto de palabras que perteneciendo a la pregunta, están muy relacionadas con el tipo de entidad de la respuesta. Si se elige adecuadamente, puede ayudar a diferenciar tipos de respuesta sin aumentar el número de tipos de pregunta. Esta utilidad puede observarse en el siguiente ejemplo:

P: *¿Cuántos animales hay en el zoo de Madrid?*

tipo de pregunta: *CUÁNTOS/AS*

entidad de la respuesta: *<cantidad>*

foco: *animal*

Con la elección de un foco se dispone de un nuevo tipo de entidad de respuesta, en este caso: *<cantidad de animales>*. Esto no significa que se haya definido esta categoría, en realidad, sólo se ha hecho una combinación. Se mantiene que la respuesta tiene que ser una cantidad, pero se añade que ésta debe pertenecer al mismo grupo que el foco.

Este trabajo puede ser más llevadero gracias a bases de datos léxicas, como **WordNet** [14], que contienen el significado de las palabras organizadas por relaciones de sinonimia, hiponimia e hiperonimia. Es decir, palabras organizadas en jerarquías de sinónimos. Como ejemplo, "*americano*" y "*estadounidense*" tendrían referencias mutuas como sinónimos (dentro de uno de sus significados), pero a la vez están referenciando conjuntamente a otro grupo de sinónimos ("*persona*", "*individuo*", "*ser*"…) en relación de hiponimia.

Es decir, estas grandes bases de datos no sólo son capaces de reconocer relaciones del tipo: *<algo1> es equivalente a <algo2>* (sinonimia) sino que reconocen relaciones como: *<algo1> es una clase de <algo2>* (hiponimia).

También se pueden utilizar sistemas (*parsers*) que analizan frases completamente, pasando por prácticamente por todos los pasos que hemos mencionado, creando un árbol con su estructura sintáctica sin perder de vista el aspecto semántico. Con ello añaden facilidad y rapidez a la hora del análisis.

Un ejemplo de analizador morfológico, reconocedor de entidades y con capacidad de POS-tagging es **FreeLing** [1], software desarrollado en la *Universitat Politècnica de Catalunya* bajo licencia del GNU[9].

Para identificar categorías semánticas existen herramientas como **IdentiFinder** [5], un algoritmo de aprendizaje que es capaz de identificar nombres y expresiones numéricas o temporales, y clasificarlos como personas, organizaciones, localizaciones, fechas...

Existen también parsers, a los que además se les entrena para que reconozcan los tipos de pregunta. Un ejemplo puede ser **CONTEX** [12] desarrollado en el *Information Sciences Institute* (University of Southern California).

2.3.3. Formación de la consulta.

Con la información que se ha obtenido en el análisis de la pregunta, se eligen las palabras que serán elegidas para la búsqueda. Por ejemplo, ante una pregunta como

"*¿Qué película ganó el oso de oro en la berlinale 2005?*"

un sistema típico formaría una consulta con una estructura similar a la siguiente:

película AND ganó AND oso AND oro AND berlinale AND 2005

En este ejemplo sólo se han elegido las palabras que contienen información y se ha formado la consulta de forma que el buscador la pueda entender. Pero, hay sistemas que, dependiendo del tipo de pregunta, hacen una consulta más personal. Así, por ejemplo, si se conoce que la respuesta debe ser *<cantidad>* incluyen las unidades como palabra a buscar. Por ejemplo, si se preguntara "*¿Cuánto mide el Everest?*" estos sistemas formarían una consulta similar a "*Everest AND mide AND metros*".

2.3.4. Expansión de la búsqueda.

Como ya se comentó anteriormente, no hay preguntas difíciles sino preguntas que no encajan en los datos de los que se dispone. Para minimizar este problema se utilizan métodos de expansión de búsqueda. Es decir, se

[9] Acrónimo recursivo de "GNU No es Unix". Es una comunidad de desarrollo de sistemas de software libre, que se financia y tiene soporte legal gracias a la FSF (*Free Software Foundation*). [Ver: http://www.gnu.org/gnu/thegnuproject.es.html]

utilizarán medios para no buscar la respuesta de una manera específica, sino con un amplio rango de posibilidades. De entre estos métodos, los más usados son los siguientes:

Expansión por sinónimos

Debido a que es posible que la solución que se encuentre no esté redactada con las mismas palabras que la pregunta, en vez de sólo buscar las palabras concretas en las que ha sido formulada la pregunta, se añaden a la búsqueda sinónimos de éstas. Para ello, se pueden utilizar bases de datos léxicas como **WordNet** (en inglés) o **EuroWordNet** [19] (en otros idiomas europeos).

Lematización

Esta expansión permitirá que el sistema sea capaz de buscar palabras independientemente de la forma en que aparezcan. En idiomas como el Inglés esto es relativamente sencillo, ya que, por ejemplo, la conjugación de sus verbos no es muy extensa. Sin embargo, si se trabaja con un idioma como el español, la expansión verbal puede resultar inviable.

Ante una lengua como la española, lo mejor es utilizar lematizadores, analizadores lingüísticos que identifican familias de palabras más allá de las raíces. Por ejemplo:

lemas: *volver / raíz*

raíces: *volv-*, *vuelv-*, *vuelt- / raíz*, *raíc-*

Un sistema con estas características es **MACO** [2], un analizador de la lengua española desarrollado por el *Laboratori de Recerca en Lingüística Computacional* de la *Universitat de Barcelona* y por el grupo de Lenguaje Natural de la *Universitat Politécnica de Catalunya*. Esta herramienta, además de identificar las distintas raíces de los verbos, es capaz de detectar otro tipo de expresiones como fechas ("*5 de Mayo de 2005*", "*2/2/02*"), nombres compuestos ("*María Elena*", "*Ministerio de Cultura*"), abreviaturas o siglas, expresiones de varias palabras ("*sin embargo*", "*por lo tanto*"), etc.

Reformulación de la pregunta

Una manera muy útil de intentar reducir distancias entre la manera en que se ha formulado la pregunta, y las posibles formas en que estén redactados los documentos, es la creación de expresiones que puedan funcionar como respuestas. Es decir, para aumentar las posibilidades de

encontrar una respuesta correcta, se crean plantillas con tipos de frases respuesta. Por ejemplo:

P: *¿Quién inventó la cerveza?*

R1: *<alguien> inventó la cerveza.*

R2: *<alguien> fue el inventor de la cerveza.*

R3: *la cerveza fue inventada por <alguien>.*

R4: *<alguien> es el padre de la cerveza.*

R5: *<alguien> recibió la patente por la cerveza.*

Estas frases, además de poder añadirse a la consulta para que sean buscadas a través de los documentos, serán usadas por el módulo de extracción de la respuesta ya que las posibles respuestas que coincidan con este tipo de frases tienen gran probabilidad de ser la correcta.

No existen sistemas exclusivamente desarrollados para hacer este tipo de trabajos, o no son completos. Por lo que será dentro de cada sistema particular dónde se implemente la creación de estas paráfrasis.

El funcionamiento de esta parte del sistema se basa en la creación de patrones. Para cada tipo de pregunta lo primero será crear uno con su formulación enunciativa, después vendrá el turno de las frases equivalentes, y por último, los patrones de dónde se podría inferir la respuesta. Una vez creados los patrones, la aplicación de éstos es sencilla, como puede verse en el siguiente ejemplo:

Pregunta: *¿Quién es el marido de <alguien1>?*

Patrón enunciativo: *"<alguien2> es el marido de <alguien1>."*

equivalente: *"<alguien2> está casado con <alguien1>."*

inferir de: *"la boda de <alguien2> y <alguien1>…"*

2.4. Búsqueda de información.

En este bloque se realiza la búsqueda del grupo de palabras que se han elegido en el módulo anterior, dentro de la colección de documentos de que se

disponga. Este tipo de búsqueda suele ser realizada por recuperadores de texto, cuyo funcionamiento, en el caso más simple, se explica a continuación:

La base principal es la *indexación* de los documentos. En la que los índices van a ser las palabras que éstos contienen. En otras palabras, se creará un archivo donde se encuentren todas las palabras contenidas en los documentos y, en cada una de estas palabras, se guardará una referencia a todos los documentos en los que aparezca.

Para realizar la indexación, primero se analizan los documentos, se extraen las palabras o grupos de palabras existentes, se ponderan por su importancia y se utilizan como índice. Generalmente habrá una lista de palabras vacías (stop-word-list), que no serán incluidas por su carencia de significado.

La **Figura 3** muestra un ejemplo cómo se realiza una indexación. En ella puede verse como se formaría un archivo índice a partir de tres archivos (doc. 1, doc. 2 y doc. 3) que contienen algunas palabras sueltas.

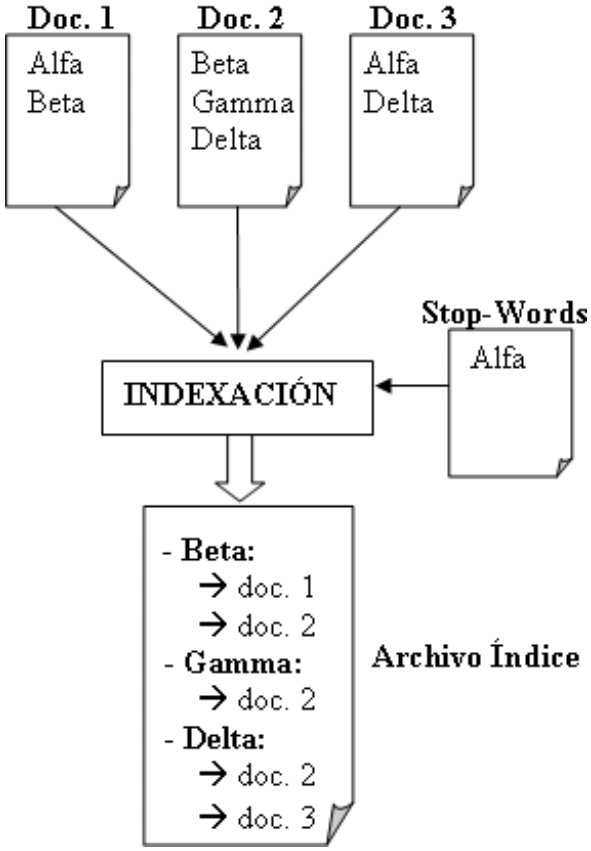

Figura 3. *Creación de un archivo índice.*

Después de la indexación, la búsqueda se realiza por comparación de las palabras en la consulta con las existentes en el archivo índice. La salida del sistema será una lista con los documentos que contienen las palabras de la consulta.

Cada sistema particular utiliza su propio modelo de RI. El más sencillo es la búsqueda *binaria*, dónde la relevancia de una palabra se mide como 1 si aparece en el documento o 0 si no lo hace. Otros modelos pueden ser, la puntuación por *frecuencia* de aparición de la palabra en el documento, o la puntuación por s*imilitud* (mediante la utilización de fórmulas heurísticas).

Siguiendo con el último ejemplo, en la **Figura 4** se muestra cual sería la respuesta de un sistema con búsqueda binaria ante una consulta simple, utilizando el archivo índice que se ha creado anteriormente.

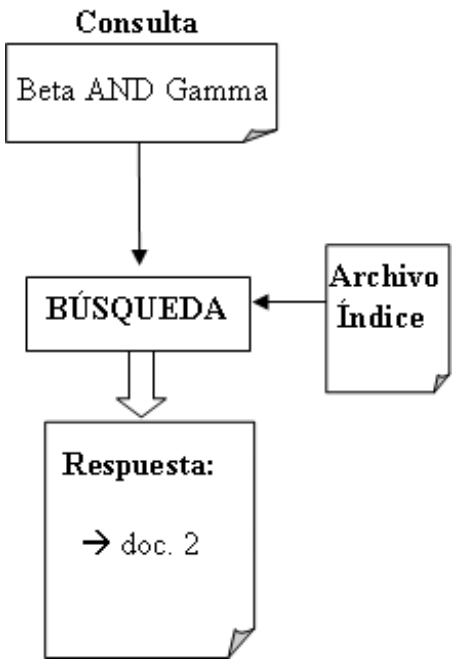

Figura 4. *Ejemplo de búsqueda.*

También es posible hacer la recuperación por párrafos (de tamaño variable o constante) o por documentos. Algunos sistemas han demostrado que la búsqueda por párrafos es más efectiva ya que elimina la dependencia del tamaño del documento, parámetro que es importante en casos como la utilización de relevancia en frecuencia.

Casos particulares de estos sistemas son:

- **MG** [17], software bajo licencia del GNU que utiliza un sistema de *diccionarios* para la indexación. Logra reducir mucho el espacio necesario comprimiendo los archivos índices en estos diccionarios, mediante codificación Huffman.

- **IR-N** [13], un buscador de la Universidad de Alicante que utiliza internamente métodos de expansión de búsqueda (sinónimos, lemas...) e incluso búsqueda bilingüe.

2.5. Extracción de la respuesta.

Una vez en este punto, se dispone de un gran conjunto de posibles soluciones a la pregunta planteada. El trabajo de este bloque del sistema es escoger, de entre todas las posibilidades, la respuesta correcta. Para realizar esto se suelen utilizar distintos modelos de puntuación que evalúan las respuestas proporcionadas por el módulo anterior. A continuación se propone un posible esquema típico para esta parte del sistema.

2.5.1. Selección de frases relevantes.

Los fragmentos de texto devueltos por el modelo de búsqueda se dividen en frases. Después, son eliminadas aquellas que no tienen ninguna relación con la pregunta planteada.

La relevancia de las frases se comprueba puntuando cada una de ellas proporcionalmente al número de palabras (o expresiones) contenidas, que coincidan con las pertenecientes a la pregunta. Las frases que no contienen ninguna expresión de la consulta son excluidas.

2.5.2. Selección frases candidatas a respuesta.

Las frases supervivientes son analizadas, identificadas sus categorías semánticas y etiquetadas. Este procesado, junto con el que se realiza en el análisis de la pregunta permite hacer un estudio conjunto de pregunta y respuesta y realizar un análisis de *concordancia*.

Lo primero es elegir las expresiones que pueden ser candidatas, mediante la comprobación del tipo *semántico*. A partir de aquí, sólo se tendrán

en cuenta aquellas que coincidan con la categoría que se dedujo en el análisis del tipo de pregunta.

Después se puntuará la posible respuesta en función del parecido a la pregunta; en función del número de palabras que coincidan, de su orden, proximidad o cualquier otro parámetro que se le pueda ocurrir al desarrollador.

Si se ha usado *reformulación*, en la expansión de la consulta, la puntuación se dará dependiendo del patrón (del peso que se le otorgara en el momento de su creación) con el que haya encajado, en el caso de que lo haya hecho con alguno.

2.5.3. Combinación de las candidatas.

Las que han pasado la anterior criba serán nuevamente analizadas. Aquellas que sean iguales o equivalentes (por ejemplo: "*un millón*" y "*1.000.000*") se unirán en una sola.

Pero a la hora de decidir cuál es la candidata elegida, el número de veces que una misma candidata se repite puede ser un dato significativo; por ello, se suele guardar una valoración proporcional a la frecuencia de aparición de la candidata con respecto al número total de ellas. Esta valoración es conocida como puntuación en *frecuencia***.**

2.5.4. Ponderación con el número de resultados.

Para disponer de más criterios a la hora de decidir cuál es la mejor respuesta, hay sistemas que realizan una búsqueda en la web para puntuar a las candidatas a respuesta. Su esquema general suele ser:

1. Formular de manera enunciativa la pregunta; tal y como se hizo en el módulo de expansión de la pregunta.

2. Rellenar la frase creada con cada una de las candidatas.

3. Buscar en Internet y obtener el número de resultados.

El número de páginas encontradas es la medida de la veracidad de la respuesta creada. Esta puntuación sólo suele usarse como complemento, ya que sólo los políticos creen que si una mentira se repite muchas veces se convierte en realidad.

2.5.5. Selección de la respuesta.

Para realizar la selección de la respuesta cada sistema utiliza una combinación distinta de las puntuaciones, referidas en este apartado o de otras, lo que lleva a la obtención de un ranking de candidatas en el que la mejor valorada será considerada la respuesta correcta.

Una manera sencilla de realizar la combinación, podría ser la multiplicación normalizada de las puntuaciones, aunque cada sistema particular, dependiendo de los criterios de diseño, busca su solución ad-hoc.

Capítulo 3.

Descripción del Sistema

"He mirado con sus ojos, he escuchado con sus oídos, y te digo que es el indicado; o por lo menos, lo más adecuado que podemos encontrar."

El Juego de Ender. **Orson Scott Card**

3.1. Introducción.

En este capítulo se mostrará cual es el esquema de funcionamiento del sistema de QA que se ha desarrollado, ***Q^{u}A-C*** *(Question Answering - Carlos III)*, cuya función principal es la de contestar, lo más correcta y ajustadamente posible, a las preguntas introducidas por los usuarios, utilizando para ello la información contenida en Wikipedia[10].

3.2. Presentación/GUI.

La interfaz de usuario, basado en PHP[11], consta de dos páginas: la de inicio, donde se introducen las preguntas y la de resultados, donde se muestran las respuestas encontradas.

Esta interfaz es muy fácil de usar y muy intuitiva, como puede verse en la **Figura 5**; consta de un campo de texto en el que se insta al usuario a introducir la pregunta que desee.

[10] Wikipedia, la enciclopedia libre. [Ver: http://es.wikipedia.org]
[11] PHP es un acrónimo de "PHP: Hypertext Preprocessor". Lenguaje de código abierto que puede ser embebido en páginas HTML (típicas páginas web) y que se ejecuta en el servidor. [Ver: http://www.php.net/]

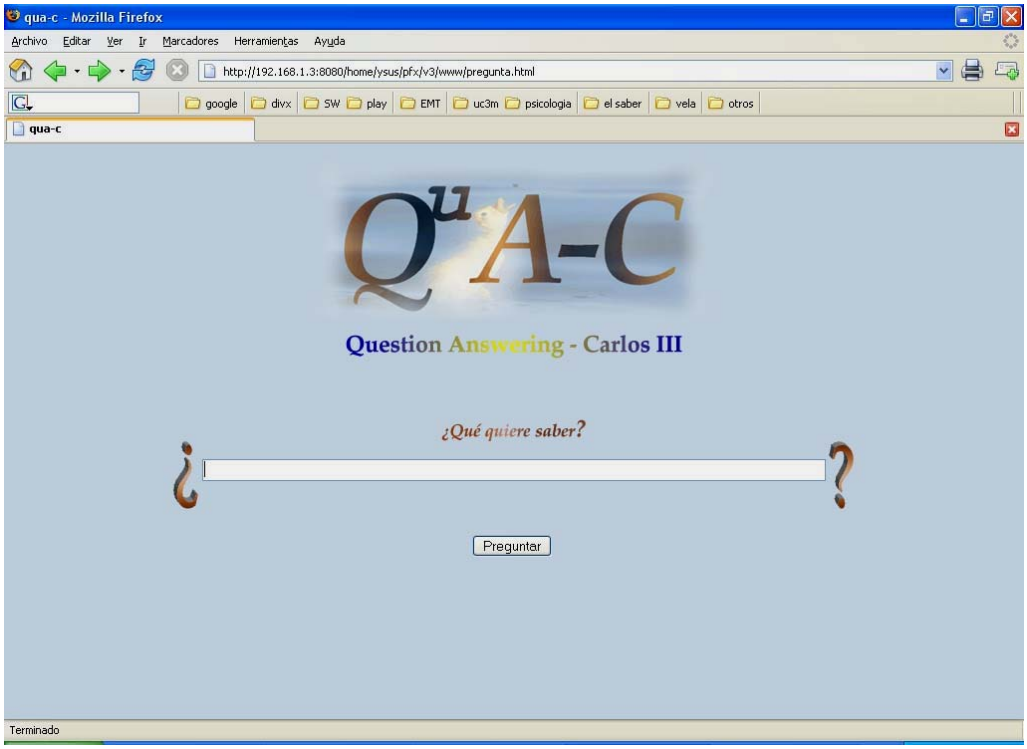

Figura 5. *Página de inicio de* ***Q^{u}A-C.***

Una vez introducida la pregunta, simplemente pulsando "Enter" o haciendo clic en el botón "Preguntar", el sistema empieza a buscar las posibles respuestas, que serán mostradas en la página de resultados, como puede verse en la **Figura 6**.

En esta página se muestran hasta cinco posibles respuestas ordenadas por relevancia. Cada respuesta consta de dos apartados:

- Una frase, sacada de los artículos de Wikipedia, que tratará de contestar a la pregunta formulada; es el primer campo de cada respuesta y se muestra entrecomillado.
- Referencia al artículo donde ha sido encontrada:
    - Nombre del artículo.
    - Nombre de la sección
    - Lugar que ocupa la oración dentro del artículo.
    - Enlace al artículo en cuestión.

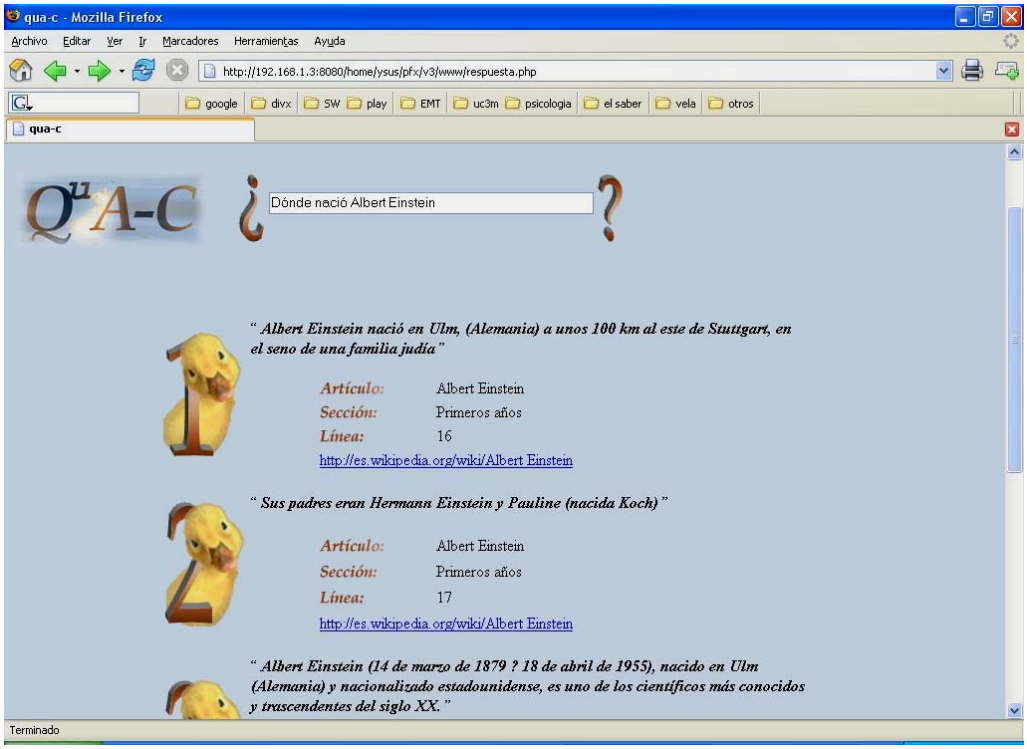

Figura 6. *Página de resultados de* ***Q^{u}A- C****.*

En esta página es posible introducir más preguntas sin necesidad de volver a la anterior, mediante el uso del cuadro de texto situado en la parte superior.

En caso de no encontrar ninguna respuesta a la pregunta, muestra un mensaje disculpándose por su ignorancia, permitiendo la formulación de una nueva pregunta o la reformulación de la anterior.

3.3. Fuente de información.

De la correcta elección de este parámetro va a depender, en buena medida, el porcentaje de aciertos de un sistema de Question Answering; ya que si la respuesta no se encuentra contenida dentro de ella, el sistema nunca podrá encontrarla. En el caso concreto de ***Q^{u}A-C***, se ha elegido Wikipedia.

3.3.1. Wikipedia.

Ésta es una enciclopedia grande y variada, continuamente actualizada y revisada, de libre acceso, bajo licencia GNU, y que es posible descargarse en varios formatos para su posterior uso. Se escribe de forma colaborativa por

voluntarios, permitiendo que, la gran mayoría de los artículos, sean modificados por cualquier persona con acceso a un navegador web. La versión en castellano consta de más de 130.000 artículos, y está en continuo crecimiento.

Fue fundada el 15 de enero de 2001 por Jimmy Wales, aunque la versión en castellano no apareció hasta mayo de ese año. Es desarrollada en el sitio web Wikipedia.org haciendo uso de su software "*wikiMedia*"[12], cuyo nombre proviene del hawaiano *Wiki Wiki*, que significa *rápido*. Actualmente, en su conjunto, dispone de más de 4.600.000 de artículos[13].

Las tres características esenciales del proyecto Wikipedia que definen conjuntamente su función en web son:

1. Es una enciclopedia, entendida como soporte que permite la recopilación, el almacenamiento y la transmisión de la información de forma estructurada.

2. Es un "wiki", por lo que, con pequeñas excepciones, puede ser editada por cualquiera.

3. Es de contenido abierto y utiliza la licencia GFDL[14].

Aunque todo esto queda resumido en la frase que utiliza su creador para describir el proyecto: "*un esfuerzo para crear y distribuir una enciclopedia libre de la más alta calidad posible a cada persona del planeta en su idioma*"[15].

3.3.2. Descarga de Wikipedia.

Una de las principales ventajas que ofrece Wikipedia, para su uso en un sistema de QA, es la posibilidad de descargar su contenido para poder así tratarlo.

Su licencia posibilita la obtención tanto de su software como de sus artículos, pudiendo encontrar instrucciones detalladas sobre como descargar las distintas versiones y formatos en su página de descargas[16].

De entre todos los archivos disponibles en castellano[17], ***Q*u*A*-*C*** utiliza un archivo xml que contiene la última versión de todos los artículos. La

[12] http://www.mediawiki.org/wiki/MediaWiki
[13] Datos ofrecidos por la propia Wikipedia. [Ver: http://en.wikipedia.org/wiki/Wikipedia]
[14] GNU Free Documentation Licence. [Ver: http://www.gnu.org/copyleft/fdl.html]
[15] http://mail.wikipedia.org/pipermail/wikipedia-l/2005-March/038102.html
[16] http://en.wikipedia.org/wiki/Wikipedia:Download
[17] http://download.wikimedia.org/eswiki/

ventaja que ofrece este formato es la facilidad y libertad para el uso, fragmentación o tratamiento de su contenido.

3.4. Arquitectura interna.

El sistema desarrollado sigue el mismo esquema general que cualquier sistema de QA que, como se vio en el capítulo anterior, es el siguiente:

1. Un usuario realiza una pregunta al sistema, en su propio idioma y de una manera natural.

2. El sistema recoge la pregunta, la analiza y elige los parámetros de la búsqueda que se realizará a continuación.

3. Mediante estos parámetros, el motor de búsqueda examina los documentos que conformen su fuente de información y devuelve los más relevantes.

4. Se dividen los documentos obtenidos y se clasifican las posibles respuestas.

5. Se muestra al usuario la respuesta más probable.

Estos puntos suelen agruparse en tres grandes bloques: análisis de la pregunta, búsqueda de información y extracción de la respuesta.

3.4.1. Pre-procesado de la información.

Como se vio en el apartado **2.4**, la base principal de los motores de búsqueda es la indexación. Por ello, si se quiere utilizar la búsqueda de información como parte del sistema, es preciso realizar la indexación de los documentos de que se disponen como paso previo. En los sistemas de QA este proceso se debe realizar también con anterioridad a todos los pasos que se han enumerado en el esquema de funcionamiento.

En el caso concreto de ***Q^{u}A-C***, se han tomado decisiones de diseño que hacen necesario un tratamiento previo de la información, anterior al indexado. Estas decisiones son:

- Utilización de la oración como unidad mínima de información.

- Lematización de la información.

- No distinción de acentos.

La oración como unidad de información

La indexación de documentos enteros dentro de los sistemas de QA, aunque muy utilizada, es una técnica que implica un tratamiento y troceado posterior de los documentos devueltos por el motor de búsqueda, para que la respuesta ofrecida al usuario sea concreta.

En el caso de ***Q^{u}A-C***, se considera que una oración es una unidad de información suficientemente concreta como para servir de respuesta pero, a la vez, con información suficiente como para ser indexada.

De esta manera, el troceado de la información se realiza con anterioridad al indexado. Ahorrándose así uno de los pasos en la extracción de la respuesta y consiguiendo, por tanto, mayor rapidez en la comunicación con el usuario.

El proceso es el siguiente:

1. Se dispone de una fuente de información consistente. En este caso, los artículos de Wikipedia.

2. Se divide cada uno de los artículos en frases.

3. Cada una de estas frases es indexada como un documento independiente.

Por lo tanto, las búsquedas que se realicen en un futuro utilizarán las frases como base, y los documentos que devuelvan dichas búsquedas no serán sino oraciones que podrán ser utilizadas directamente como respuestas.

Lematización de información y distinción de acentos

Las ventajas que ofrece el uso de la lematización en la "expansión de la búsqueda" durante el Análisis de la pregunta (ver apartado **2.3.4.**), son bien conocidas. Sin embargo, para dotar al sistema con esta capacidad es necesario introducir un bloque lematizador en dos puntos del esquema:

- Antes del indexado, toda la información que se indexe debe estar previamente lematizada.

- Después de la introducción de la pregunta, la pregunta debe ser lematizada para proceder a su análisis.

Con respecto a la distinción de palabras acentuadas, no existe un consenso sobre qué es más beneficioso. Para aclarar las cosas, un sistema que utilice distinción entre palabras acentuadas (con tilde) y no acentuadas (sin tilde) sería incapaz de devolver un documento en el que apareciera la palabra "*matemáticas*", si se ha utilizado como parámetro de búsqueda la palabra "*matematicas*"; los sistemas que no realizan esta distinción, interpretan de la misma manera cualquiera de las dos palabras.

Durante el proceso de desarrollo de ***Q^{u}A-C***, se realizaron pruebas con ambos modelos de sistemas. En general, se observó que las personas (entre las que se incluye el propio autor) tienden a olvidar alguna que otra tilde, por lo que se llegó a la conclusión de que era más beneficioso el no utilizar esta distinción.

3.4.2. Contestando preguntas.

Después del bloque anterior, la información está "cargada" en el sistema, por lo que ***Q^{u}A-C*** está listo para contestar preguntas. Su esquema de funcionamiento, como recordará, está dividido en tres grandes bloques: análisis de la pregunta, búsqueda de información y extracción de la respuesta.

Análisis de la pregunta

El objetivo de este bloque es la formación de una consulta entendible por el buscador, a partir de la pregunta introducida por el usuario.

Para ello, ***Q^{u}A-C*** realiza en primer lugar una segmentación de la pregunta para obtener sus palabras; después, lematiza éstas de la misma manera que se utilizó en la indexación de fuente de información. Una vez dispone del conjunto de lemas, correspondientes a las palabras que contenía la pregunta, elabora una lista (también lematizada) de sus sinónimos para, al final, combinarlos con los originales para formular la consulta que será la entrada del siguiente bloque.

Además de esta función, este bloque también tiene la capacidad de realizar una pequeña detección del tipo de pregunta.

Búsqueda de Información

El motor de búsqueda del sistema, utilizando la consulta que se ha creado en el bloque anterior, examinará los documentos en busca de las palabras indicadas.

Las características especiales de esta búsqueda son:

- La utilización de un archivo de palabras sin significado (stop-words).
- El reconocimiento de números.
- El reconocimiento de términos multipalabra (sin embargo, no obstante…) y de nombres (nombres propios, ciudades…).
- Irrelevancia de los acentos (tildes).
- La oración como unidad de información.

Extracción de la respuesta

En primer lugar, las oraciones devueltas por el sistema de búsqueda están lematizadas, sin acentos y sacadas de su contexto; por ello, el primer cometido es encontrar la frase original y el contexto del que fue sacada (su artículo). Después, se realiza una revaloración de las respuestas mejor colocadas, en función de su correspondencia con el tipo de pregunta que se haya detectado.

Al final, las respuestas mejor valoradas, serán mostradas al usuario; pero, antes de presentarle los resultados, será necesario dar el formato adecuado a los datos, de manera que pueda ser entendible por la interfaz de usuario.

Capítulo 4.

Adquisición del Conocimiento

CIENCIA: una manera de descubrir cosas y hacerlas funcionar. La Ciencia explica lo que sucede a nuestro alrededor en todo momento. Lo mismo hace la RELIGIÓN, pero la Ciencia es mejor porque, cuando se equivoca, ofrece unas excusas más comprensibles.
Hay mucha más Ciencia de lo que uno se imagina.
De Una Enciclopedia científica para el joven gnomo curioso, *por Angelo de Mercería.*

La Nave. **Terry Pratchett**

4.1. Introducción.

La fuente de información en la que se basa ***Q^{u}A-C*** para contestar a las preguntas que realice el usuario es, como ya se dijo, Wikipedia. El contenido de los artículos de esta enciclopedia será la base del conocimiento del sistema, pero para que este conocimiento forme parte de él, la información debe ser indexada.

Para realizar un correcto indexado de la información es necesario primero separar la información real, de la información relativa al formato que esta contenga. Con este objetivo se ha creado un sistema específico para esta fuente, que es capaz de filtrar la información relativa a los artículos, dividirla en oraciones y desechar el resto; después, estas oraciones se lematizan y se filtran los acentos, para dar paso a la indexación final. El esquema general de este proceso puede verse en la **Figura 7**.

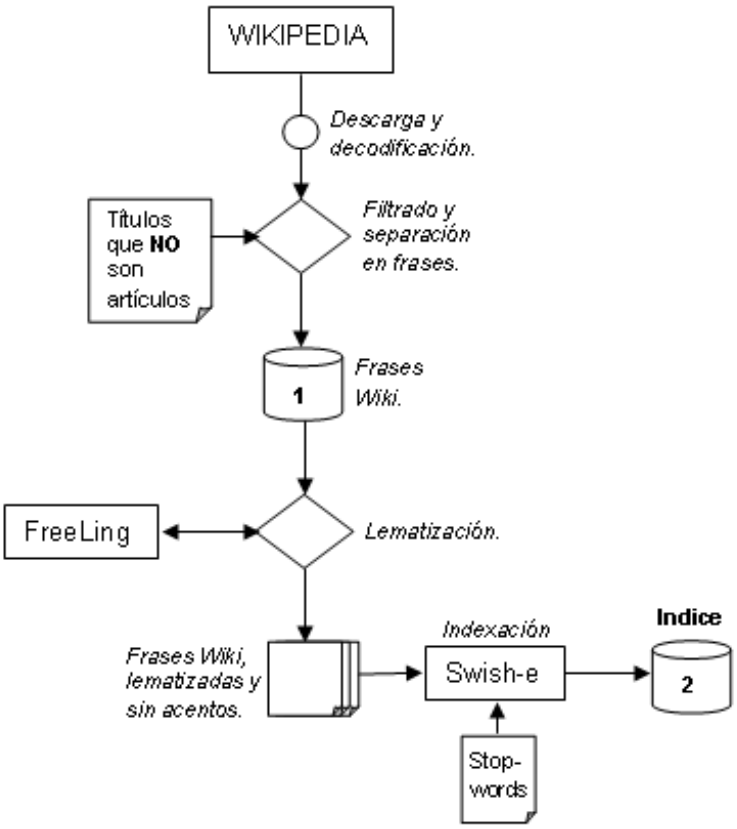

Figura 7. *Esquema de adquisición del conocimiento en* ***Q^{u}A-C****.*

En este capítulo se explicarán, en un primer momento, todas las peculiaridades relativas al formato de la información, que sean consideradas relevantes, y el modo en que se filtran, para después describir los pasos que realiza el sistema durante la lematización y la indexación; a su vez, se explicarán las dos herramientas de software libre que ***Q^{u}A-C*** utiliza durante estos procesos: **FreeLing** y **Swish-e**.

4.2. Formato de la información.

Una vez elegida la fuente de información, en este caso Wikipedia, es necesario descargar sus artículos para proceder a su tratamiento e indexado. De entre los formatos que ofrece Wikipedia como posibilidades de descarga (ver apartado **3.3.2**), se ha elegido xml, por la facilidad que ofrece este formato a la hora de tratar los datos; en concreto, el archivo que contiene las versiones más recientes (se actualiza periódicamente) de todos los artículos publicados en esta enciclopedia.

4.2.1. Formato del archivo.

En este archivo, codificado en utf8[18], los artículos de Wikipedia están dispuestos, uno detrás de otro, según el siguiente formato:

```
<page>
  <title>[…]</title>
  <id> […] </id>
  <revision>
    <id> […] </id>
    <timestamp> […] </timestamp>
    <contributor>
      <username> […] </username>
      <id> […] </id>
    </contributor>
    <comment> […] </comment>
    <text>
        […]
    </text>
  </revision>
</page>
```

Figura 8. *Formato de artículo en Wikipedia.*

Donde las distintas etiquetas:

<page> </page>, delimitan a cada artículo.

- <title> </title>, contienen el título del artículo.
- <id> </id>, es el número del artículo.
- <revision> </revision>, contienen los datos de la última revisión, a saber:
    - <id> </id>, el número de revisión.
    - <timestamp> </timestamp>, es la fecha de la última revisión.
    - <contributor> </contributor>, contiene datos del último editor del artículo:
        - <username> </username>, nombre.
        - <id> </id>, número de usuario.

[18] Utf8: 8-bit Unicode Transformation Format. [Ver: http://www.unicode.org/]

- <*comment*> </*comment*>, comentarios informativos sobre el artículo, estado de evolución, posible plagio…

- <*text*> </*text*>, contienen el cuerpo del artículo.

4.2.2. Formato del cuerpo de artículo.

Para dar formato a sus artículos, Wikipedia pone a disposición de sus autores dos mecanismos distintos: por un lado, se pueden editar directamente en formato HTML, aunque su recomendación es no hacerlo; por otro, se ofrece un formato específico, creado para la edición interna de artículos, facilitando su uso mediante páginas de edición y plantillas.

Siguiendo este formato, un artículo debe estar dividido en secciones, facilitando así su visión global y permitiendo a los lectores ver el esquema de su contenido (si el artículo está dividido correctamente en secciones, se crea una tabla índice automáticamente al principio del artículo). El principio de cada una de las secciones se distingue por estar rodeado su título por caracteres '='. El lugar que ocupe la sección dentro de la jerarquía de la página vendrá dado por el número de "iguales" que rodeen su título. Es decir, las secciones con más iguales serán subsecciones de la sección con un número de iguales menor inmediatamente anterior. Ejemplo:

```
=Sección 1=
==Sección 1.1==
=Sección 2=
==Sección 2.1==
===Sección 2.1.1===
===Sección 2.1.2===
==Sección 2.2==
```

La información básica de formato de la letra (negrita, cursiva...) se encuentra en el número de comillas simples (') que rodean a la palabra o frase; así, por ejemplo, una línea o palabra rodeada por dos comillas simples a cada lado aparecerá en cursiva, y si son tres aparecerá en negrita.

'''AC/DC''' → **AC/DC**

''compact'' → *compact*

Las listas también se pueden distinguir de manera sencilla. Se forman simplemente con la utilización de un carácter delante de la palabra o frase que se pretende incluir en la lista. Por ejemplo, el carácter '#' se usa para las listas ordenadas y el carácter '*' para las no ordenadas.

*'''Paint it Black''' (''Rolling Stones'') → • **Paint it Black** (*Rolling Stones*)

Otro elemento habitual dentro de los artículos de Wikipedia son los enlaces internos. Éstos son los encargados de comunicar unos artículos con otros, simplemente pulsando en una palabra clave. Se caracterizan por el uso de dobles caracteres de corchete ('[' y ']'); dentro de éstos aparecerá, por un parte, la palabra o frase que se mostrará en la página y por otra, la dirección del enlace a la que referencia, separados por el carácter '|'. Su estructura es la siguiente:

[[información a mostrar | dirección del enlace]]

Las tablas son otro elemento muy utilizado en la redacción de artículos. Su formato se distingue por estar contenido entre los siguientes caracteres '{|' y '|}'; dentro de estos se usan '|' para separar celdas, '- - - -' para separar filas, etc. Un ejemplo sencillo puede ser:

```
{|
!Encabezado 1!!Encabezado 2!!Encabezado 3
|----------------
|celda 1||celda 2||celda 3
|----------------
|celda 1||celda2||celda 3
|}
```

Con éstos, se formaría una tabla de 3x3 donde su primera fila es de encabezado. Las palabras que aparecen en el ejemplo son indicativas de dónde debería ir el contenido tanto de los encabezados como de las celdas, no ofreciendo ninguna información de formato.

También existen otra multitud de elementos de formato[19], pero que por ser menos utilizados en los artículos, no se van a ser expuestos aquí.

4.3. Filtrado de la información.

La información de que dispone el sistema es, por tanto, un único archivo que contiene todos los artículos de Wikipedia redactados en un formato que se ha visto en el apartado anterior.

El primer paso es decodificar este fichero, originalmente en formato utf8, y pasarlo a formato ASCII de 8 bits (ISO-8859-1); lo que permite tener un fichero de información que nuestro sistema es capaz de tratar.

Para que esta información sea indexable se creó un sistema capaz de filtrar la información real de los artículos y dividirla en oraciones, eliminando

[19] Ver: http://es.wikipedia.org/wiki/Wikipedia:Cómo_se_edita_una_página

toda la información relativa al formato. La estructura general del programa es la siguiente:

1. Selección de artículos.
2. División y selección de oraciones.
3. Filtrado de elementos de la oración.

Donde se distinguen tres etapas distintas de filtrado, a distintos niveles; de más grande a más pequeño, Artículo -> Oración -> Elemento.

4.3.1. Selección de artículos.

Dentro del archivo con los contenidos de Wikipedia, además de las últimas versiones de todos sus artículos se encuentran multitud de otros artículos: Discusión, Ayuda, Plantillas, Usuarios, etc. Éstos, aunque siguen el mismo formato que los artículos enciclopédicos, no deben ser incluidos como información relevante dentro del sistema.

Cada artículo será separado del resto por estar delimitado por las etiquetas <*page*> </*page*>, aunque, de momento, solo será tenido en cuenta su título (entre las etiquetas <*title*> </*title*>), donde se realizará el primer filtrado.

El formato que siguen los títulos de los artículos internos de Wikipedia es conocido: todos ellos contienen el carácter ':' sin espacios a sus lados, ni a izquierda ni a derecha. Ejemplo:

Wikipedia:Acerca de

Aunque este criterio, por sí solo, no puede ser utilizado como filtro de artículos, ya que existen artículos enciclopédicos que lo cumplen; por ello, se creó una lista basada en un análisis de todos los artículos que cumplían este criterio, donde aparecen todas las palabras que, situadas delante del carácter ':' indican con toda seguridad que se trata de un artículo interno. La lista de estas palabras se muestra en la **Tabla 2.**

Tabla 2. *Artículos no enciclopédicos.*

Ayuda	*Plantillas*
Ayuda Discusión	*Portada*
Categoría	*Portal*
Categoría Discusión	*Portal Discusión*
Discusión	*Transwiki*
Enciclopedia	*Usuario*
Imagen	*Usuario Discusión*
Imagen Discusión	*Wikipedia*
MediaWiki	*Wikipedia Discusión*
MediaWiki Discusión	*Wikiportal*
Modelo	*Wikiproyecto*
Plantilla	*Wikiproyecto Discusión*
Plantilla Discusión	

Además de este filtrado a nivel de artículo, existe otro filtrado de artículo que se realiza a nivel de oración; se trata del filtrado de artículos redirigidos, estos artículos se descubren al leer la primera frase, que suele empezar por "*#redirect*...", y son ignorados inmediatamente.

4.3.2. División y selección de oraciones.

Una vez realizado el filtro de artículos por título, se procede a analizar el contenido de cada uno de ellos, separándolo en oraciones. El cuerpo de cada artículo se encuentra contenido entre las etiquetas <*text*> </*text*>, el resto de la información (excepto el título) es desechado.

Este texto se separa en frases, teniendo especial cuidado en las secciones del artículo; cada oración se guardará junto con una referencia al título del artículo, sección a la que pertenece y posición que ocupa (número de frase dentro del artículo).

También dentro del cuerpo del artículo, existe mucha información de carácter interno, o que al menos, no es importante para un sistema de QA; en concreto existen multitud de mensajes que Wikipedia introduce aquí para informar al usuario del estado del artículo que está consultando, entre ellos:

- El artículo es sólo un esbozo.
- Existe información relacionada con el artículo en Wikimedia, Wikiquote o cualquier otro Wiki.
- Algún tipo de error.
- No neutralidad.
- y un largo etcétera.

La mayor parte de este tipo de información es localizada por el sistema de procesado y no pasa a formar parte de la base de datos, gracias a que suelen compartir un formato común, como empezar y terminar por dobles llaves '{{...}}'.

También hay secciones que no aportan ningún tipo de dato interesante al sistema, aunque sean parte de la información útil del artículo cuando uno lo está leyendo, y no son tenidas en cuenta; estas secciones son:

- **Véase también**, sección que ofrece referencias a otros artículos de la propia Wikipedia.
- **Enlaces Externos**, donde se encuentran las referencias a páginas útiles si se quiere seguir investigando sobre el tema.
- **Referencias**, donde se muestra la bibliografía que el autor ha usado para redactar el artículo.

Por la misma razón que en el caso anterior, la falta de información, tampoco se tratan las líneas dedicadas a informar de la categoría a la que pertenece el artículo ni las que muestran las versiones que existen del artículo en otros idiomas.

La manera de descubrir estos fragmentos sin información suele ser analizar con cuidado las líneas que consisten sólo en un enlace interno (rodeados por dobles corchetes) y comprobar si siguen el formato de artículo interno de Wikipedia (conjunto de palabras que contienen el carácter ':' sin espacios a los lados de éste).

Por último, aunque sí contienen información útil para el sistema, no se tratan los elementos de tipo "tabla". Esto es debido a la dificultad que conlleva el analizar la estructura de estos elementos y mantener la concordancia debida tanto de fila, columna, tabla y sección a la hora de guardar la información. Sin embargo, no se descarta que el sistema pudiera mejorar con el ingreso de esta información en su base de datos, por lo que se deja en consideración para trabajo futuro.

4.3.3. Filtrado de elementos de la oración.

Dentro de cada oración también puede haber elementos (caracteres) que contengan información de formato que no hará sino entorpecer las labores de búsqueda si no es eliminada.

Los principales elementos de este tipo que elimina el sistema son:

- La información relativa a los enlaces internos. Se eliminan por una parte, los corchetes, y por otra, la información relativa a la dirección del enlace; únicamente se conserva la parte que se muestra al usuario que consultaba el artículo en la web.

- La información relativa al formato de letra. Se eliminan todas las comillas indicativas de formato (negrita, cursiva...).

- Las etiquetas y símbolos HTML. Se eliminan todos los indicadores de formato escritos en este lenguaje.

Después de tener las frases filtradas de información de formato, se guarda cada una de ellas en un archivo, junto con la información relativa al título, sección a la que pertenece y posición que ocupa esa oración dentro de artículo.

4.3.4. Ejemplo.

Como todo este proceso puede resultar confuso, se va a realizar una pequeña demostración de cuáles son los pasos que se dan, con un ejemplo.

A continuación se muestra un pequeño fragmento, correspondiente a un artículo, del archivo descargado de Wikipedia:

```
  <page>
    <title>Rubielos de Mora</title>
    <id>147781</id>
    <revision>
      <id>1703592</id>
      <timestamp>2005-11-25T03:08:00Z</timestamp>
      <contributor>
        <username>LeonardoRob0t</username>
        <id>25375</id>
      </contributor>
      <minor />
      <comment>robot Añadido: it, pt</comment>
      <text xml:space="preserve">'''Rubielos de Mora''' es una localidad de la comarca [[Gúdar-Javalambre]] en la [[Provincia de Teruel]] ([[España]]) situada a unos 55 km de la capital de provincia. Se accede por la carretera comarcal C-232.

== Datos geográficos ==

Superficie:63,7 km²

Altura: 929 m

Población: 620 habitantes en 1996

Gentilicio: rubielano

== Demografía ==

Evolución de la población ( año (habitantes)):
* 1900 (2.257)
* 1910 (2.235)
* 1920 (1.863)
* 1930 (1.771)
```

** 1940 (1.362)*
** 1950 (1.268)*
** 1960 (1.196)*
** 1960 (1.196)*
** 1970 (930)*
** 1981 (666)*
** 1991 (570)*

== Monumentos ==

El pueblo cuenta con un casco urbano antiguo bastante bien conservado. Destacan una torre de la antigua muralla, el ayuntamiento renacentista del siglo XVI, varias ermitas y casas de la nobleza como la de los condes de Florida etc.

Juntos con la vecina Mora de Rubielos es un importante atractivo para el turismo local.

== Fiestas ==

**14 de septiembre (La Santa Cruz)*
**16 de julio (La Virgen del Carmen)*
**12 de octubre (La Virgen del Pilar)*

== Enlaces externos ==

**[http://www.hernandezrabal.com/espana/aragon/teruel/rubielosmora.htm] (fotos e información adicional)*
**[http://www.caiaragon.com/es/municipios/index.asp?idloc=381&tipo=0 Ficha de la población]*

[[Categoría:Localidades de Teruel]]
{{esbozo de|geografía de Teruel}}

[[it:Rubielos de Mora]]
[[pt:Rubielos de Mora]]</text>
</revision>
</page>

Figura 9. *Ejemplo de artículo de Wikipedia.*

En primer lugar, se puede ver claramente como los artículos están delimitados por las etiquetas *<page> </page>*, de esta forma están ordenados, uno detrás de otro. También se puede comprobar que la única información de interés para el sistema está contenida en las etiquetas *<title> </title>* y *<text> </text>*.

A la hora de analizarlo, lo primero es comprobar el título, en este caso *<title>Rubielos de Mora</title>* no tiene nada que haga pensar que no es un artículo enciclopédico, por lo que se proseguirá con su análisis. El siguiente paso es analizar las frases, se coge la primera:

'''Rubielos de Mora''' es una localidad de la comarca [[Gúdar-Javalambre]] en la [[Provincia de Teruel]] ([[España]]) situada a unos 55 km de la capital de provincia.

Esta frase no cumple ninguna de las condiciones para ser desechada, por lo que pasa a ser guardada, pasando antes por el filtrado de elementos. El resultado sería el siguiente:

Rubielos de Mora es una localidad de la comarca Gúdar-Javalambre en la Provincia de Teruel (España) situada a unos 55 km de la capital de provincia.
Rubielos de Mora

Donde puede verse que la información de formato es eliminada y que es añadida información relativa a su procedencia (cuando la oración pertenece a la sección principal, no se añade información sobre ella).

Este proceso se seguirá realizando de la misma manera a lo largo del artículo; cogiendo otra línea al azar, por ejemplo: *Altura: 929 m*, el archivo que se generará será el siguiente:

Altura: 929 m
Rubielos de Mora: Datos Geográficos

Y el proceso seguirá hasta que llegue a la sección *Enlaces externos*, que el sistema no procesará. Después, ya no procesará ninguna de las líneas por cumplir los criterios que se especificaron en la sección **4.3.2**:

- *[[Categoría:Localidades de Teruel]]*, por ser un enlace que coincide con artículos internos Wikipedia.
- *{{esbozo de|geografía de Teruel}}*, información para el usuario de Wikipedia, el artículo es sólo un esbozo.
- *[[it:Rubielos de Mora]]* y siguiente, indican que ésta página existe en otros idiomas, en este caso, italiano. Es rechazada por la misma razón que el primer punto.

En la última línea se puede leer la etiqueta *</text>* por lo que el sistema dejará de procesar texto hasta que se encuentre con el título del artículo siguiente.

4.4. Tratamientos pre-indexación.

En este punto, la información es totalmente indexable. Se dispone de una colección de documentos que contienen las posibles respuestas a las preguntas que realice el usuario; por lo que toda esta información se podría indexar ya en el motor de búsqueda (cargar la información en el sistema) que se use en la búsqueda de información, **Swish-e** (ver apartado **4.6.1**). Sin embargo, en este caso, la información debe pasar todavía por dos procesos:

- Lematización.
- Eliminación de acentos.

4.4.1. Lematización.

La lematización es un proceso que se usa como método de expansión de la búsqueda (ver apartado **2.3.4**), consiste en reducir las palabras a sus lemas elementales, consiguiendo identificar familias de palabras y

considerándolas como una sola lo que es una gran ventaja a la hora de buscar información; por ejemplo, las búsquedas serán independientes del tiempo verbal utilizado tanto en la pregunta como en la respuesta.

Pero para poder disfrutar de esta ventaja a la hora de realizar la búsqueda, primero es necesario lematizar la información antes de indexarla, para lo que se ha se ha utilizado **FreeLing**[20] (ver apartado **4.5**).

Mediante un programa creado para interactuar con esta librería, todos los archivos creados en el apartado anterior (que contienen las oraciones de los artículos de Wikipedia) se lematizan.

El esquema general de funcionamiento de este programa, basado en C++ es el siguiente:

1. Recogida del texto de entrada.
2. Segmentación de sus elementos.
3. Agrupamiento en frases de los elementos segmentados.
4. Análisis morfológico de las frases.
5. Post-análisis y etiquetado.
6. Selección y devolución de frases lematizadas.

4.4.2. Eliminación de acentos.

Otra de las expansiones de la búsqueda que se ha introducido es la no distinción de acentos (ver apartado **3.4.1**); por ejemplo, para ***Q^{u}A-C***, será lo mismo buscar la palabra "terapéutico" que "terapeutico".

No obstante, al igual que en el caso de la lematización, es necesario hacer un tratamiento a la información, antes de que ésta sea indexada. En este caso, el tratamiento es simple, se eliminan las tildes de las palabras que van a ser indexadas si están acentuadas gráficamente.

Esta tarea debe hacerse en último lugar, justo antes de la indexación, porque de otra manera podría acarrear algún error, dado que hay palabras que, dependiendo de si llevan o no tilde, pueden pertenece a familias distintas; si se quitan los acentos antes de la lematización, es posible que Freeling devuelva un lema equivocado.

[20] http://garraf.epsevg.upc.es/freeling/

4.4.3. Ejemplo.

Una vez separada toda la información de los artículos en líneas y está limpia de información de formato, se procede a lematizar y quitar los acentos. Volviendo al ejemplo de la **Figura 9**, ante el archivo con la frase:

Rubielos de Mora es una localidad de la comarca Gúdar-Javalambre en la Provincia de Teruel (España) situada a unos 55 km de la capital de provincia.
Rubielos de Mora

El sistema devolvería la siguiente salida:

rubielos_de_mora ser uno localidad de el comarca gudar-javalambre en el provincia de teruel (españa) situar a uno 55 km de el capital de provincia.
rubielos_de_mora

4.5. Freeling.

Freeling [1] es una herramienta de análisis del lenguaje de código abierto, bajo licencia LGPL[21] desarrollado por el Centro TALP (*Centre de Tecnologies i Aplicacions del Llenguatge y la Parla*)[22] de la Universidad Politécnica de Cataluña, y que ha contado con la colaboración, en el desarrollo de diccionarios y gramáticas, del CLiC (*Centre de Llenguatge i Computació*)[23] de la Universidad de Barcelona.

Esta herramienta está constituida por una librería que provee de servicios de análisis del lenguaje a cualquier aplicación que lo necesite; no es por tanto un software independiente, sino un conjunto de funciones y diccionarios, accesibles desde cualquier aplicación.

Las características que ofrece esta librería son:

- Segmentación.
- Agrupamiento en frases.
- Análisis morfológico.
- Detección de nombres propios (por ejemplo, de ciudades).

[21] Lesser General Public Licence. [Ver: http://www.fsf.org/licensing/licenses/lgpl.html]
[22] http://www.talp.upc.edu/talp/
[23] http://clic.fil.ub.es/

- Detección de fechas y números.
- Etiquetado (cada palabra con su categoría).
- Análisis sintáctico de la frase, superficial o por dependencias.
- Posibilidad de uso en distintas lenguas: español (castellano), catalán, gallego, inglés e Italiano.

El diccionario español contiene más de 81.000 formas distintas, correspondientes a más de 7.000 lemas; a lo que hay que añadir un potente módulo de análisis de sufijos que permite la detección de los pronombres en las formas verbales (por ejemplo: "*morirse*", "*dímelo*"...) y multitud de diminutivos y aumentativos en sustantivos y adjetivos. Todo esto le permite clasificar correctamente, según sus propios datos, más del 80% de los términos de un texto de carácter general.

El diccionario semántico está basado en EuroWordNet, y si es necesario se facilita la inclusión completa de esta herramienta, bajo sus términos de licencia.

4.5.1. Segmentación y agrupamiento en frases.

La segmentación es una parte sencilla pero fundamental de cualquier análisis lingüístico; consiste en separar cada oración en sus elementos más simples, las palabras, para poder tratarlas como elementos independientes. Así, cada palabra tendrá entidad propia y será posible su tratamiento mediante métodos más complejos; es el primer paso para el análisis morfológico.

Además, para que los análisis se realicen en el contexto adecuado, después de la segmentación (que ha separado todo el texto en elementos simples) Freeling agrupa en frases estos elementos (*sentence splitting*).

4.5.2. Análisis morfológico y etiquetado.

Este proceso recibe, como entrada, un texto segmentado y agrupado en frases, sobre él se realizaran varios procesados secuencialmente para obtener el análisis morfológico:

1. El primero es el detector de locuciones (combinación de palabras que funcionan como una sola), por ejemplo: no obstante, de rechupete, una vez que, alguno que otro...

2. Búsqueda en diccionarios y detección de sufijos.

3. Detección de números, es capaz de interpretar de la misma manera las cantidades, independientemente de cómo estén escritas. Por ejemplo, 100.000, 100000 o cien mil, serán para Freeling la misma cosa.

4. Detección de fechas, es capaz de interpretar fechas en distinto formato, como 06/06/06 y 6 de Junio de 2006.

5. Detección de nombres propios, de personas, ciudades...

6. Cálculo de probabilidades y manejo de palabras desconocidas, donde se calculará, mediante fórmulas heurísticas, la probabilidad de que una palabra pertenezca a una categoría u otra por medio del contexto y serán tratadas las palabras desconocidas.

La manera de describir las distintas categorías a las que puede pertenecer una palabra son descritas mediante el uso de unas etiquetas basadas en las creadas por el grupo EAGLES (*Expert Advisory Group on Language Engineering Standards*)[24] de las que se muestra, a continuación, un pequeño resumen.

Introducción a las Etiquetas Morfológicas[25]

Estas etiquetas están formadas por un conjunto de cifras y letras que en las que se codifica toda la información de la palabra analizada. Su categoría gramatical, género, número, persona, tiempo, modo…

En la **Tabla 3** se muestra una lista de las etiquetas utilizadas, aunque sólo hará referencia al primer carácter de estas etiquetas, que es el indicativo de la categoría gramatical y, en algún caso, al segundo, el que informa del tipo dentro de la categoría.

[24] http://www.ilc.cnr.it/EAGLES96/home.html
[25] La guía completa se puede encontrar en:
http://garraf.epsevg.upc.es/freeling/doc/userman/parole-es.html

Tabla 3. *Etiquetas de Freeling.*

A...	*Adjetivo*
R...	*Adverbio*
D...	*Determinante*
NC...	*Nombre Común*
NP...	*Nombre Propio*
V...	*Verbo*
P...	*Pronombre*
C...	*Conjunción*
I	*Interjección*
SP...	*Preposición*
F...	*Signo de Puntuación*
Z	*Cifra*
W	*Fecha u Hora*

A continuación, para entender mejor el análisis morfológico de Freeling, se muestra un ejemplo sacado de la demostración, disponible en Internet[26], de esta herramienta; se presenta el siguiente texto de entrada:

> *El gato come pescado y bebe agua.*
> *Mi amigo Juan Mesa se mesa la barba.*

Después de pasar por los procesos de segmentación y splitting, realiza el análisis morfológico, obteniéndose el resultado mostrado en la **Figura 10** y **Figura 11**:

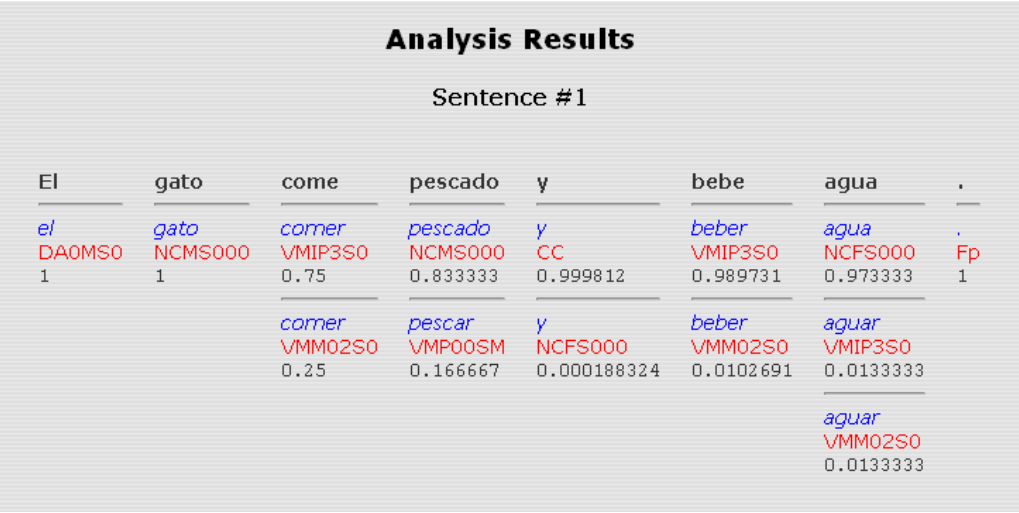

Figura 10. *Análisis morfológico de la frase 1.*

[26] http://garraf.epsevg.upc.es/freeling/demo.php

Figura 11. *Análisis morfológico de la frase 2.*

Como se puede comprobar, las frases son diferenciadas (*splitting*) y los elementos son tratados individualmente (segmentado). Por lo que el análisis morfológico es distinto para cada frase.

También se puede apreciar cómo calcula, basándose en el contexto, las probabilidades correspondientes a las distintas categorías posibles. En la primera frase, esto puede apreciarse perfectamente en "*pescado*", ya que puede pertenecer a dos categorías distintas: puede ser un verbo, en cuyo caso su lema sería "*pescar*"; o puede ser un nombre común, en cuyo caso sería directamente "*pescado*"; después del cálculo de probabilidades, la mejor colocada es el nombre común.

Un ejemplo muy similar es el caso de "*amigo*", en la segunda oración. En este caso, la palabra puede pertenecer a tres categorías distintas (verbo, adjetivo o nombre común), pero con sólo dos lemas diferentes ("*amigar*" y "*amigo*"); no obstante, el análisis de probabilidades aclara la situación otorgando un 86% a la posibilidad del nombre común.

Otro de los casos que merece la pena resaltar es la diferenciación que realiza Freeling de las dos ocurrencias de la palabra mesa en la segunda frase: en el primer caso, formando una sola palabra junto con Juan, otorgándole la categoría de nombre propio; sin embargo, la segunda vez que aparece esta palabra la interpreta, o bien como nombre común, o como verbo.

Desambiguación

A esta etapa, le corresponde el turno de elegir, entre todas las categorías a las que puede pertenecer, la etiqueta que mejor se ajusta a este caso.

Basándose en el análisis morfológico realizado, el sistema realiza un reprocesado de la oración, utilizando fórmulas heurísticas basadas en los algoritmos de Brants [6]. Después de este segundo análisis, más exhaustivo que el anterior, selecciona una única categoría para cada palabra, la que tiene una mayor concordancia tanto con el contexto semántico como con la posición correspondiente dentro de la oración.

Utilizando las frases anteriores, los resultados serían los mostrados en la **Figura 12.**

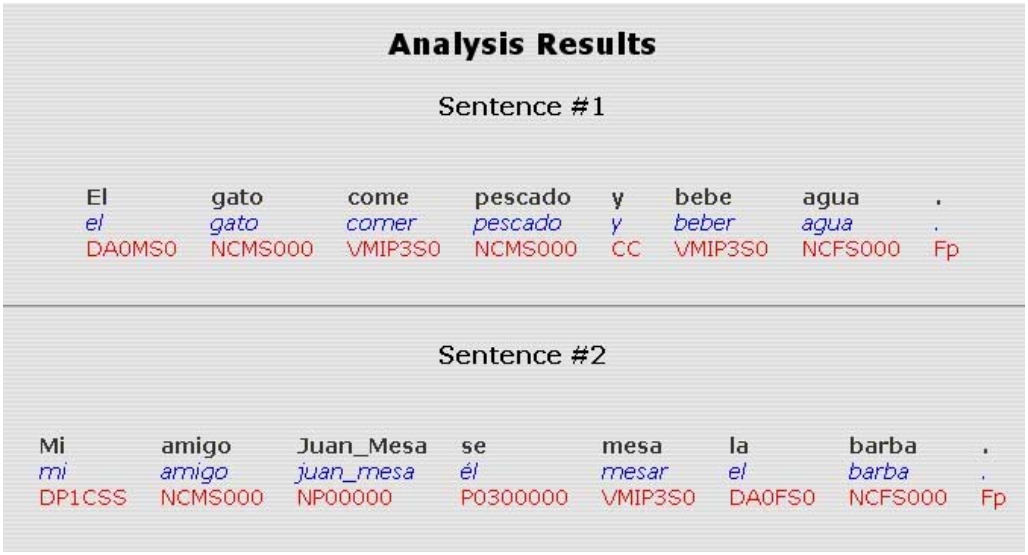

Figura 12. *Ejemplo de desambiguación en Freeling.*

El caso más significativo de este ejemplo es la palabra "*mesa*". En el anterior análisis se remarcó la diferencia en el tratamiento entre las dos palabras mesa de la segunda oración, para la segunda de ellas (en minúscula), como recordará, se ofrecían distintas posibilidades. Si vuelve a mirar ese ejemplo, podrá comprobar que se atribuía mayor probabilidad a la opción de que su lema fuera "*mesa*", perteneciendo a la categoría de nombre común; sin embargo, tras realizar este análisis, es elegido como lema "*mesar*", al ser etiquetada la palabra como verbo principal de la oración (*VM…*).

Es, por tanto, mediante la conjunción de estos dos análisis como se obtiene un nivel de aciertos mayor en la elección de los lemas, no debiendo utilizar únicamente el análisis morfológico a la hora de analizar una frase, porque su cálculo de probabilidades no es del todo significativo.

4.5.3. Análisis sintáctico.

Después de los dos análisis anteriores, el texto está preparado para servir de entrada a los analizadores sintácticos; a partir de este paso, los lemas elegidos no cambiarán sino que se realizará un análisis en un nivel superior.

Este tipo de análisis agrupa sintácticamente elementos de la frase, realizando búsquedas de dependencias entre ellos. Esto resulta muy útil en módulos de traducción automática entre idiomas en los que haya gran divergencia lingüística, como en el caso del castellano y el euskera.

Freeling se basa en dos analizadores principalmente: TXALA [3], un analizador libre de dependencias para el castellano y TACAT [4], una herramienta de análisis sobre texto etiquetado.

Este tipo de herramientas sobrepasa el análisis necesario en un sistema QA monolingüe, por lo que no se ahondará más en el tema. Sin embargo, a continuación, se muestra un ejemplo que aclarará el tipo de análisis al que se está haciendo referencia.

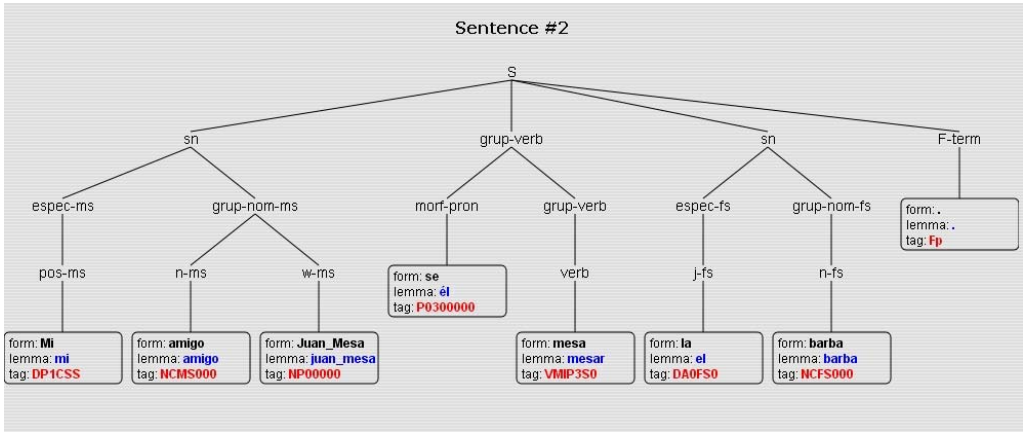

Figura 13. *Ejemplo de análisis sintáctico en Freeling.*

Utilizando como ejemplo la segunda frase del apartado anterior, puede verse como se divide la frase en dos sintagmas nominales y un grupo verbal, cada uno de ellos formado por distintos elementos.

4.6. Indexación.

Después de todos estos tratamientos (filtrado de información, separación en frases, lematización…) la información está lista para ser

indexada. El encargado será **Swish-e**[27], un sistema de indexado, gratuito (bajo licencia GNU), rápido, flexible y bien valorado según artículos en revistas informáticas como *LinuxJournal.com* [15].

Cada frase, como documento independiente (lematizada y sin acentos), pasará a formar parte de la base de datos del sistema mediante la indexación; el sistema de indexado creará un archivo índice en el que aparecerán todas las palabras que aparezcan en esos archivos (excepto las palabras vacías o stop-words) guardando en cada una de ellas, una referencia a todos los documentos en los que aparecen.

4.6.1. Swish-e.

Swish-e (*Simple Web Indexing System for Humans – Enhanced*) es un motor de búsqueda, rápido, fácil de usar, y de código abierto. Fue creado en un principio por Kevin Hughes en 1994, sólo como herramienta de indexado de páginas Web. Después, fue transferida a la Universidad de California (UC Berkeley), donde se desarrolló como proyecto para la creación de su biblioteca digital[28], con *Sun Microsystems.* Ahora, bajo licencia GPL, su código sigue bajo desarrollo en *SourceForge.net*[29].

Swish-e es capaz de indexar un gran número de documentos en diferentes formatos, desde el texto plano, hasta formatos como PDF o PostScript; su principal ventaja consiste en la rapidez, tanto en el indexado como en la búsqueda, con bajos consumos de memoria; es altamente configurable e integrable con otras herramientas y, además, está diseñado "para humanos", por lo que su manejo y configuración es sencillo.

Permite la inclusión de expresiones regulares como reglas de inclusión o exclusión de documentos en el indexado, incluyendo el tratamiento de etiquetas (tipo XML) como propiedades de éstos. También ofrece distintos formatos de salida, con las listas de los documentos más relevantes y limitar la búsqueda a determinados tipos de archivos, entre otras características.

Indexado

El cometido principal del indexado es la creación de un archivo índice donde aparezcan todas las palabras que se puedan encontrar en los documentos.

---

[27] http://swish-e.org/
[28] http://sunsite.berkeley.edu/
[29] Web donde se desarrollan más de 100.000 proyectos de software libre, la más importante del mundo de estas características. [Ver: http://sourceforge.net/]

Para la construcción de este índice, es necesario un archivo de configuración; en él, además de los parámetros necesarios para la indexación, se deben incluir los tipos de archivos que se desean indexar.

El primer paso es separar en palabras, interpretando como tales conjuntos de caracteres alfanuméricos. En el caso del castellano, es necesario que introduzcamos en el archivo de configuración cuáles son estos caracteres, debido a que, por defecto, el carácter 'ñ' no es interpretado como tal.

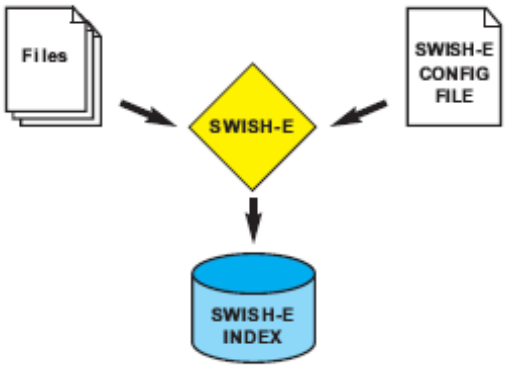

Figura 14. *Indexación en Swish-e.*

Es posible, además, realizar un indexado "difuso" (*fuzzy*), basado en algoritmos que detectan las raíces de las palabras; pero, esta capacidad está en vías de desarrollo y la versión disponible es muy pobre. En el caso de ***Q^{u}A-C***, este tipo de procesado no sería necesario, ya que se incorpora como paso anterior la lematización.

Búsqueda

Las palabras introducidas en la consulta serán buscadas a lo largo del archivo índice, para devolver después una lista con los documentos más relevantes.

La búsqueda se puede realizar sobre palabras, frases u otros elementos, como palabras truncadas. Las palabras truncadas se forman añadiendo el carácter '*' detrás de una palabra, con lo que se buscarán todos los términos que empiecen por los caracteres de esa palabra (e.g. "*biblio**" → "*biblioteca*", "*bibliografía*"…).

La puntuación que se le dará a los documentos será de tipo binario, valorando por igual todos los elementos introducidos en la búsqueda, independientemente del orden.

Swish-e y QA.

En el campo del QA, Swish-e ha sido muy bien valorado por estudios como el que realizó la Universidad de Groningen en 2004 [18]. En este estudio se analizaban las capacidades de siete buscadores en el contexto de QA, estas aplicaciones (Zettair[30], Xapian[31], Zebra[32], Lucene[33], Amberfish[34] Managing Gigabytes (MG)[35] y Swish-e) fueron sometidas a distintas pruebas, en las que era medido el nivel de acierto.

Como método de evaluación se propusieron varias medidas dependiendo de la longitud del documento devuelto (archivo, párrafo o frase), que fueron evaluadas según dos criterios: *doc MRR*, porcentaje de casos en los que la respuesta se encontraba entre los tres documentos mejor clasificados; *answer MRR*, porcentaje de casos en los que el documento mejor clasificado contenía la respuesta.

En los resultados, que se muestran en la **Tabla 4**, se puede comprobar que Swish-e fue uno de los que mejor resultado obtuvo en este estudio, consiguiendo ser el mejor clasificado según el segundo criterio.

Tabla 4. *Puntuaciones obtenidas en el estudio.*

	documents		paragraphs		sentences	
MRR (in %)	doc	answer	doc	answer	doc	answer
Swish-e	26.02	54.01	28.62	43.52	23.85	32.87
Zettair	32.10	52.69	29.90	42.09	28.32	31.04
Xapian	28.25	50.49	30.11	41.41	25.14	28.90
Zebra	26.50	45.06	27.79	37.53	25.47	30.67
Lucene	29.74	47.87	30.14	36.48	27.82	29.61
Amberfish	21.05	44.31	20.67	28.05	21.15	23.06
MG	20.86	39.98	20.98	22.53	21.18	15.44

4.6.2. Stop-Words.

Existen, en todas las lenguas, un conjunto de palabras que por su excesiva repetición ofrecen muy poca información para aplicaciones de búsqueda. Estas palabras (determinantes, conjunciones, preposiciones,

[30] http://www.seg.rmit.edu.au/zettair/
[31] http://www.xapian.org/
[32] http://www.indexdata.dk/zebra/
[33] http://lucene.apache.org/java/docs/
[34] http://www.etymon.com/tr.html
[35] http://www.cs.mu.oz.au/mg/

artículos y demás), definidas usualmente como "vacías", suelen ser eliminadas antes del indexado para no distorsionar las búsquedas posteriores.

Por ejemplo, se debe indexar la frase: "*Un día vino un hombre con un sombrero*". La palabra "*un*" aparece tres veces y, sin embargo, no aporta ninguna información. Si este tipo de palabras no se eliminan, ante una pregunta como "*¿Qué es un planeta?*", la oración anterior dará un resultado de tres coincidencias (tres veces "*un*"), mayor que el que daría: "*Planeta, cuerpo masivo que orbita una estrella y que no posee brillo propio*" (una coincidencia, "*planeta*").

Para evitar este tipo de situaciones, los motores de búsqueda incluyen listas de estas palabras para no procesarlas ni en el indexado ni durante la búsqueda; en el caso de Swish-e, esta lista no está disponible en castellano, pero permite la posibilidad de introducir una externa, mediante el archivo de configuración.

La lista utilizada[36] por ***Q^{u}A-C*** ha sido desarrollada por Lluís Padró[37], miembro del *Departamento de Lenguaje y Sistemas Informáticos* de la Universidad Politécnica de Cataluña y uno de los creadores de Freeling (ver [1]).

[36] La versión completa de este archivo, formada por 184 palabras, se puede consultar en: http://www.lsi.upc.es/~padro/freqs/empty.sp.2.gz
[37] http://www.lsi.upc.es/~padro/

Capítulo 5.

Contestando Preguntas.

"Y es que la máquina podía tener en cuenta perfiles psicológicos, compatibilidad de ADN y otros factores […] Pero lo que nunca sería capaz de procesar y comprender es que a ella le salía un gazpacho riquísimo […], ni lo que esto significaba verdaderamente."

Izlup y Sorbina. **José Tomas Romero**

5.1. Introducción.

El objetivo de este capítulo es describir cuál es el esquema de funcionamiento de ***Q^{u}A-C*** desde la introducción de una pregunta en el sistema hasta que se muestran las posibles respuestas. Este esquema, como recordará, está dividido en tres grandes bloques (análisis de la pregunta, búsqueda de información y extracción de la respuesta) que se unen de manera secuencial.

El primer bloque es el encargado del tratamiento de la pregunta y su conversión en una consulta entendible por el motor de búsqueda, que se ocupará, en el segundo bloque, de buscar las posibles respuestas existentes en la fuente de información, para finalizar escogiendo en el tercer bloque cuales de éstas tienen más posibilidades de ser la respuesta correcta y mostrarla al usuario.

Aunque durante la redacción de este capítulo se mantendrá esta estructuración en bloques, debido a su valor didáctico, en el esquema general que se muestra en la **Figura 15** se ha prescindido de ella para conseguir una mayor claridad.

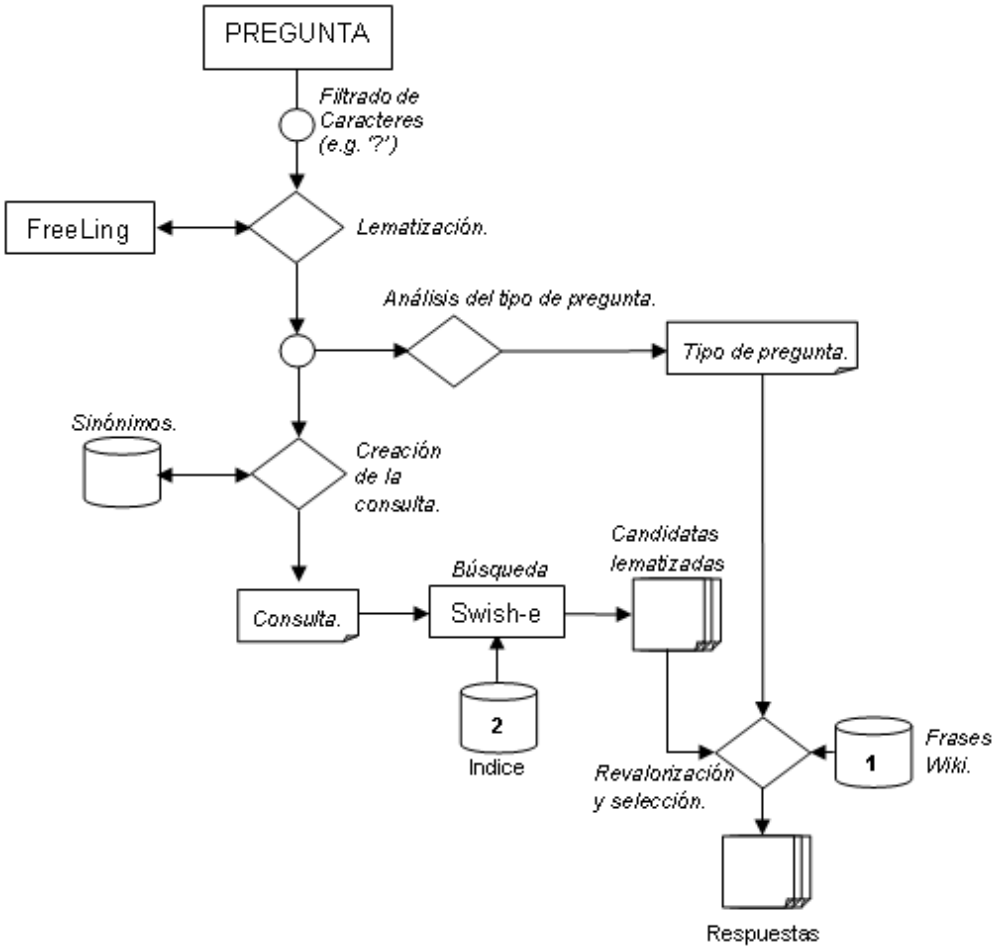

Figura 15. *Esquema de cómo **Q"A-C** contesta una pregunta.*

5.2. Análisis de la pregunta.

En primer lugar, la pregunta se segmenta, se obtienen sus palabras y, después, se lematizan; todo ello, con el mismo proceso que se utilizó durante el tratamiento pre-indexación, basado en la librería Freeling.

Antes de formular la consulta, se realiza una expansión por sinónimos y se detecta de manera simple el tipo de pregunta, para finalizar redactando una consulta en el formato específico del buscador.

5.2.1. Lematización.

Después de que el usuario introduce una pregunta, el primer paso antes de crear la consulta es su lematización. De esta manera, al obtener los

lemas en un primer momento, otros procesos que vendrán después, como la expansión por sinónimos se verán beneficiados por él.

El proceso se realiza por el mismo programa, pero esta vez, en vez de analizar artículos de Wikipedia, analiza la pregunta introducida por el usuario. Al igual que en el caso anterior, se devolverá una frase, equivalente a la de entrada, pero únicamente formada por lemas.

En este caso, la eliminación de las tildes no se realizará en este punto; después se realizará la expansión por sinónimos, por lo que es conveniente mantener todavía la información relativa a las tildes, para evitar posibles confusiones, por ejemplo, para no confundir "*éste*" (pronombre) con "*este*" (punto cardinal), a la hora de añadir sinónimos a la búsqueda.

5.2.2. Expansión por sinónimos.

El mecanismo principal en el que se basa un sistema de QA para contestar es buscar las palabras que aparecen en la pregunta dentro de una colección de documentos, para devolver un fragmento de información en relación con ellas. La probabilidad de acierto se basa entonces en la correspondencia que exista entre la manera de redactar la pregunta y los documentos en los que se realice la búsqueda.

Para aumentar esta probabilidad, se utilizan métodos de expansión de la búsqueda, como la lematización (vista en el apartado anterior) o la expansión por sinónimos, donde además de buscar las palabras que aparecen en la formulación de la pregunta, se buscan sus sinónimos.

OpenThesaurus

OpenThesaurus[38] es un conjunto de proyectos desarrollados como software libre cuyo objetivo es crear diccionarios de sinónimos accesibles gratuitamente y compatibles con programas de procesado de texto como OpenOffice[39]. La versión de este proyecto en castellano, OpenThesaurus-es[40] (bajo licencia LGPL), contiene más de 20.000 palabras y está en continuo desarrollo.

Esta versión consta principalmente de dos archivos, uno de índices y otro con la base de datos de sinónimos. No contiene ningún tipo de software ya que está ideado como complemento del procesador de textos de

[38] http://sourceforge.net/projects/openthesaurus/
[39] Grupo de desarrollo de herramientas de "oficina" (procesador de texto, hoja de cálculo, editor de presentaciones...) multiplataforma y de código abierto. [Ver: http://es.openoffice.org/]
[40] http://openoffice-es.sourceforge.net/thesaurus/

OpenOffice, pero el formato de los archivos es sencillo, por lo que se puede diseñar fácilmente una aplicación que devuelva los sinónimos de una palabra introducida.

Las palabras de las que se pueden obtener sinónimos están colocadas al principio de una línea, después, separado por un carácter '|', una cifra que indica el número de líneas contiguas que contienen sinónimos de esa palabra; en las siguientes líneas (tantas como indique el número anterior), que empezarán por el carácter '-', se encuentran los sinónimos, separados por el carácter '|'.

A continuación, se muestra un fragmento de este archivo:

```
abismo|4
-|atolladero|bancal|barranca|barranco|cauce|descolgadero (NoRAE)
-|acantilado|cuesta|declive|escabrosidad|escarpa|escarpadura|precipicio
-|averno|báratro|infierno
-| despeñadero|precipicio|quebrada|rambla|ramblazo|rehoyo|sima|talud
abitaque|1
-|amarradero|bita|poste
abjuración|2
-|apostasía|refractación
-|negación|traición
```

Figura 16. *Ejemplo de archivo de sinónimos.*

Esta colección de sinónimos está creada bajo la protección de SourceForge.net como un proyecto conjunto de multitud de personas, que van aportando sus conocimientos y su tiempo desinteresadamente; por lo tanto, no se trata de un documento completamente contrastado ni ofrece ninguna garantía. Sin embargo, sí existe algún tipo de control, por ejemplo: en él aparecen tanto palabras aceptadas por la RAE[41] como no aceptadas, pero suele estar indicado en la propia palabra, como se puede ver en la palabra "*descolgadero*", en la primera línea de sinónimos de "*abismo*".

5.2.3. Sinónimos en *Q^uA-C.*

Para poder utilizar este tipo de expansión de la búsqueda, lo primero que se necesita un diccionario de sinónimos, en formato electrónico y de acceso gratuito. Este tipo de diccionarios no abundan en castellano, aunque existen algunos, entre ellos el de El Mundo[42] (utilizado por el autor en la redacción de este documento) y el de la Universidad de Oviedo[43]. El problema que presentan estos diccionarios es que sólo son accesibles vía Web.

[41] Real Academia Española. [Ver: http://www.rae.es/]
[42] http://www.elmundo.es/diccionarios/
[43] http://tradu.scig.uniovi.es/sinon.html

Para ganar rapidez y fiabilidad, se decidió utilizar un diccionario que pudiera descargarse y utilizar en local, ésta es la razón principal por la que se utiliza OpenThesaurus-es.

Para poder utilizar sus archivos se ha desarrollado una aplicación basada en PHP capaz de devolver una lista de los sinónimos de una palabra introducida. Después del proceso de lematizado de la pregunta se buscarán los sinónimos de los lemas correspondientes mediante este programa. Estos procesos deben realizarse en este orden, aunque pueda parecer más lógico el contrario.

La principal razón para utilizar este orden es que ofrece la posibilidad de encontrar sinónimos del verbo principal de la oración. El diccionario de sinónimos utilizado -en realidad, cualquier diccionario- sólo contiene verbos en infinitivo, sin embargo, en la mayor parte de las oraciones (a menos que se hable "en indio") los verbos aparecen conjugados, por lo que sería imposible encontrar sinónimos de éstos sin una lematización previa.

5.2.4. Análisis del "tipo de pregunta".

En castellano existen diversos tipos de interrogativos, que se pueden clasificar, según [10], de la siguiente forma:

- Determinativos interrogativos, grupo formado por "*qué*", "*cuánto*", "*cuántos*", "*cuánta*", "*cuántas*", antepuestos a un sustantivo al que "actualizan", por ejemplo: "¿*Qué libro* has leído?" o "¿*Cuánto dinero* ganas?".

- Pronombres interrogativos, en el que se incluyen los anteriores, si no llevan sustantivo (por ejemplo: "*¿Qué lees?*" "*¿Cuánto ganas?*") y "*quién*", "*quiénes*", "*cuál*" y "*cuáles*".

- Adverbios interrogativos, formado por "*dónde*", "*cuándo*", "*cuánto*" y "*cómo*".

Pero el número de categorías que se pueden formar a partir de ellos es mucho más elevado, y para poder realizar un análisis exhaustivo del tipo de pregunta se necesitaría una cantidad de tiempo equivalente o mayor a la del proyecto que se ha realizado.

Sin embargo, se ha querido incluir en $\boldsymbol{Q^uA\text{-}C}$ un pequeño módulo que detecte algún tipo de pregunta. Con este objetivo, se diseñó un módulo que, dependiendo del tipo de pregunta, clasificara la categoría de la respuesta uno de estos tres grupos:

- Número, se espera que la respuesta sea un número ante las preguntas del tipo: "*cuándo*", "*cuánto/a/os/as*" o "*en qué año/mes/día/fecha*".

- Nombre propio, la respuesta pertenecerá a esta categoría cuando se realice una pregunta del tipo: "*dónde*" o "*quién*" y aquellas que utilicen la fórmula "*en qué país/ciudad*"; excepto que, tratándose de una pregunta de tipo "*quién*", vaya seguida del verbo "*ser*" y un nombre propio (e.g. "*¿Quién es Joaquín Sabina?*").

- Cualquiera, en caso de no encajar en ninguna de las categorías anteriores no se hace suposición alguna sobre la categoría de la respuesta.

5.2.5. Construcción de la consulta.

El paso final del bloque de análisis de la pregunta es la construcción de una consulta que contenga las palabras que quieren buscarse en el siguiente módulo, búsqueda de información.

Estas palabras son de dos tipos, las que formaban parte de la pregunta y los sinónimos de éstas. Para controlar el grado de importancia que se le da a cada una de estas palabras dentro de la búsqueda, se pueden utilizar dos tipos de "conectores":

- *AND*, para indicar máxima importancia. Es decir, si una consulta consiste en un grupo de palabras unidas por este conector, en la búsqueda sólo serán considerados relevantes los documentos en los que aparezcan todas las palabras incluidas en dicha consulta.

- *OR*, para indicar importancia relativa. Es decir, ante una consulta realizada sólo con este conector, serían considerados relevantes los documentos en el momento en el que apareciera cualquiera (una o varias) de las palabras.

El objetivo de la consulta será, por tanto, buscar los documentos en los que aparezcan todas las palabras de la pregunta, pudiendo sustituir alguna de ellas por un sinónimo, en caso de que no aparezca.

Para que la búsqueda sea independiente de los acentos, al igual que se hizo antes de la lematización, se realiza un filtrado de éstos antes de formular la consulta.

En un caso concreto, por ejemplo, ante la pregunta: "*¿Quién descubrió el ácido acetilsalicílico?*", se formaría la consulta que se muestra en la **Figura 17**.

```
(descubrir AND acido AND acetilsalicilico)
 OR
((descubrir OR (acertar OR adivinar OR atinar OR descifrar OR encontrar OR
                hallar OR resolver OR crear OR idear OR imaginar OR
                inventar OR trazar OR alumbrar OR revelar OR comunicar)
AND
(acido OR (afilado OR agudo OR caustico OR duro OR hiriente OR mordaz
                OR ofensivo OR acre OR agrio OR intransigente)
AND
(acetilsalicilico))
```

Figura 17. *Ejemplo de consulta.*

5.3. Búsqueda de información.

A partir de la consulta construida en el bloque anterior, se realizará una búsqueda a través de los documentos, utilizando para ello el motor de búsqueda de Swish-e.

En esta etapa ***Q^{u}A-C***, se encargará de manejar esta herramienta, configurar los parámetros correspondientes e indicar cuáles van a ser las palabras que no se buscarán (stop-word-list).

5.4. Extracción de la respuesta.

En este bloque se recogen las respuestas mejor valoradas por el motor de búsqueda y se analizan para crear un nuevo ranking de favoritas, en el que no sólo se tenga en cuenta el número de palabras coincidentes.

Hay que tener en cuenta también que las candidatas a respuesta que devuelve el motor son frases lematizadas, que no deben mostrarse al usuario, por lo que tendrán que buscarse las frases originales, filtradas de información de formato, que se crearon durante la indexación.

Los pasos que se dan en este bloque para mostrar al usuario las respuestas mejor valoradas son:

1. En un primer momento, se lee el fichero que devuelve Swish-e como resultado de la búsqueda y se toman en consideración las quince mejores respuestas.

2. Después, se buscan los equivalentes sin lematizar de estas frases.

3. Para cada una de ellas, se pondera la puntuación recibida en el motor de búsqueda por la longitud de la frase, dando así un valor añadido a las respuestas concretas.

4. Las frases originales se analizan morfológicamente para buscar dentro de ellas, con ayuda de Freeling, la categoría a la que debe pertenecer la respuesta. Las frases que contengan la categoría correcta serán mejor valoradas y adelantarán, en el ranking de candidatas, a las que no la contengan.

5. Por último, se muestran los cinco mejores resultados.

Capítulo 6.

Pruebas

"Entonces comienza el ejercicio práctico. Me temo que la lección teórica me ha agotado demasiado. En efecto, estoy desfallecido pero esto forma parte de mi sino. Estiro como puedo la mano y agarro la botella; la descorcho temblando. [...] Repito: no deseaba emular a los hombres; los emulaba en busca de una salida."

Informe para una academia. **Franz Kafka**

6.1. Introducción.

Los sistemas de QA deben ser evaluados por medio de una colección de preguntas suficientemente amplia y variada, sólo de este modo los resultados podrán ser considerados relevantes. En el caso concreto de ***Q^{u}A-C***, se ha elegido el conjunto formado por doscientas preguntas utilizado en la evaluación de los sistemas de QA presentados en CLEF 2006.

A lo largo de este capítulo se expondrán cuáles han sido los resultados obtenidos por el sistema, tras someterlo a la batería de preguntas, para estudiar después cuál ha sido la influencia concreta de dos de los módulos empleados en la expansión de la búsqueda: expansión por sinónimos y análisis del tipo de pregunta.

6.2. Las preguntas.

El conjunto de preguntas utilizado, conseguido gracias a César de Pablo, ha sido creado con el objetivo de evaluar sistemas de QA en CLEF (la versión completa puede verse en el **Apéndice A**), lo que aporta fiabilidad a los

resultados obtenidos. Estas preguntas versan sobre los más diversos temas, están formuladas de maneras muy variadas, exigen distintos niveles de conocimiento e incluso incluyen "preguntas trampa" (como la 153). A modo de ejemplo, a continuación se muestran algunas de ellas:

31: "*¿Quién descubrió el cometa Shoemaker-Levy?*"

60: "*¿Qué es el tóner?*"

118: "*¿Qué robaba el oso Yogi?*"

130: "*¿Qué organismo presidía Primo Nebiolo durante los Campeonatos del Mundo de atletismo de Gotemburgo?*"

153: "*¿Cuántas veces ha ganado Zinedine Zidane el US Open?*"

165: "*¿Qué países forman el Consejo de Cooperación del Golfo?*"

175: "*¿Cuándo se independizó Surinam?*"

Figura 18. *Ejemplo de preguntas CLEF 2006.*

6.3. Las respuestas.

Las respuestas que $\boldsymbol{Q^uA\text{-}C}$ devuelve al usuario son, como ya sabe, frases sacadas de los artículos de Wikipedia, por lo tanto, en la mayoría de los casos la respuesta ofrecida no es una contestación directa a la pregunta realizada sino que se deducirá de esta.

Durante la evaluación del sistema se han clasificado las respuestas en cinco categorías:

- **Respuesta correcta**, cuando la frase devuelta por $\boldsymbol{Q^uA\text{-}C}$ en primera posición incluya, de manera clara, la respuesta correcta. Un ejemplo de esta categoría es la pregunta 25: "*¿Cómo se le llama también al Síndrome de Down?*", cuya respuesta se muestra en la **Figura 19**.

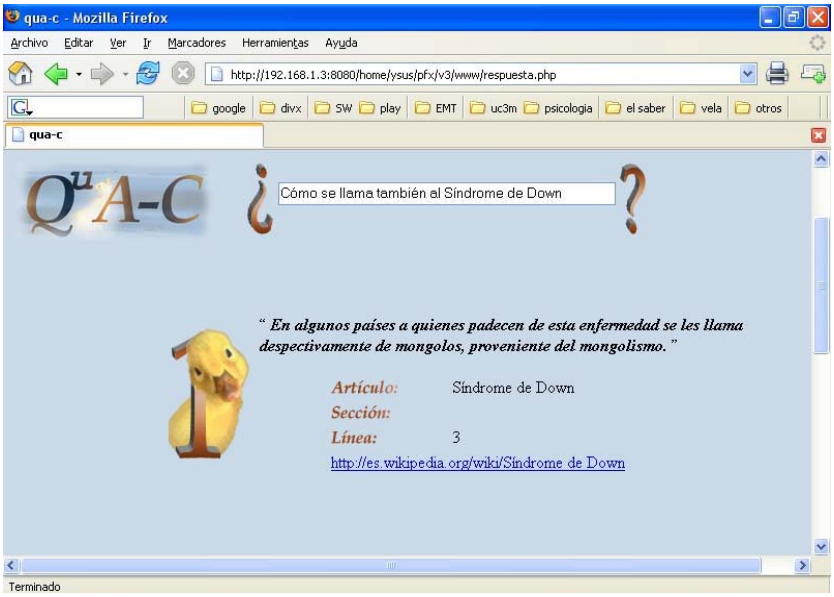

Figura 19. *Ejemplo de Respuesta correcta (pregunta 25).*

- **Respuesta en 5**, cuando en alguna de las cinco frases devueltas por ***Q^uA-C*** se encuentre la respuesta correcta (este grupo contiene al anterior). Como ejemplo se muestra, en la **Figura 20**, la respuesta a la pregunta 154: "*¿Quién es el astronauta que ha estado más tiempo en el espacio?*"

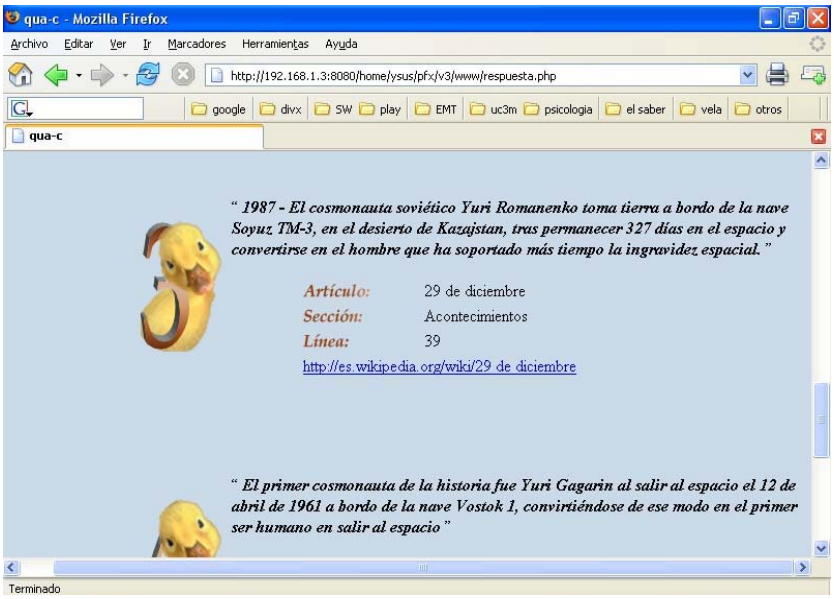

Figura 20. *Ejemplo de Respuesta en 5 (pregunta 154).*

- **Aproximada**, cuando la respuesta no se encuentre de manera clara en ninguna de las frases devueltas, pero se pueda deducir de alguna de ellas. Por ejemplo, la pregunta 29: "*¿Qué altura tiene la*

Torre Eiffel?" a la que el sistema responde como puede verse en la **Figura 21**.

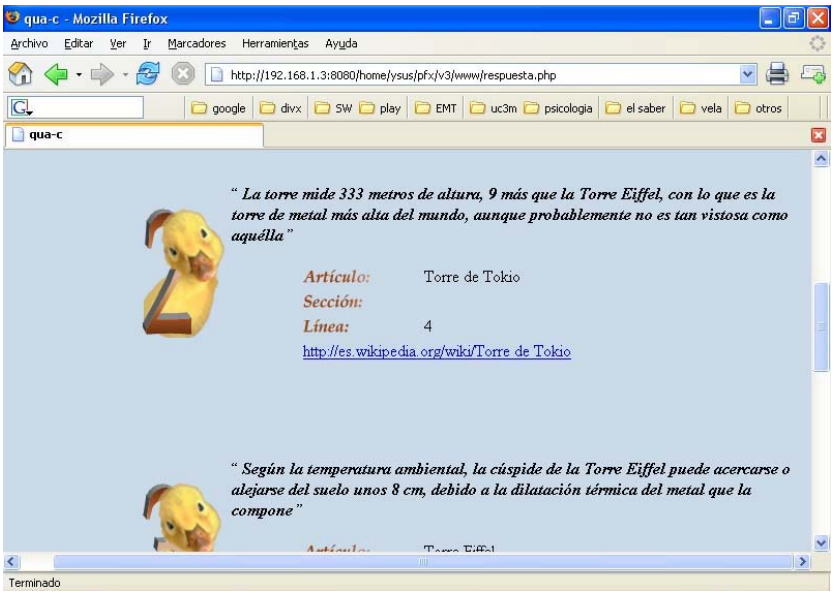

Figura 21. *Ejemplo de respuesta aproximada (pregunta 29).*

- **Fallo**, cuando ninguna de las oraciones devueltas contenga la respuesta, ni correcta ni aproximada.

- **No Respuesta**, cuando el sistema no responde. Un ejemplo se muestra en la **Figura 22**, ante la pregunta 186: *"¿En qué ciudad de Zelanda pasaba varias semanas al año Jan Toorop entre 1903 y 1924?"*

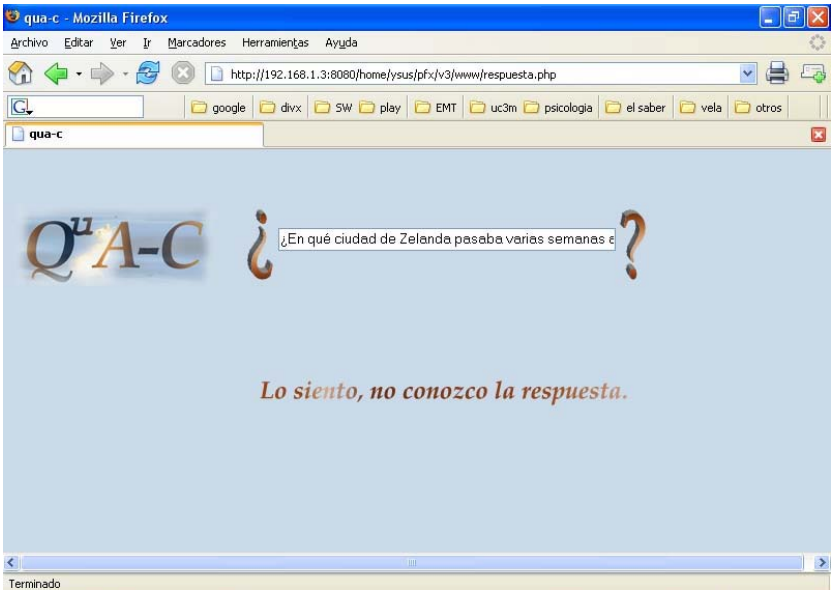

Figura 22. *Ejemplo de No Respuesta (pregunta 186).*

6.4. Evaluación de Q^uA-C.

En este apartado se analizarán los resultados obtenidos por el sistema en tres de sus implementaciones, cada vez más complejas: básico, con expansión por sinónimos y, por último, con análisis del tipo de pregunta.

6.4.1. Q^uA-C básico.

Esta versión del sistema tiene todas las capacidades expuestas a lo largo de la memoria excepto el módulo de expansión por sinónimos y el de análisis del tipo de respuesta. Es, por tanto, un sistema de QA completamente funcional.

Tras someter a este sistema a la colección completa de las preguntas mencionadas anteriormente, los resultados fueron los siguientes:

Tabla 5. *Resultados de Q^uA-C básico.*

Respuesta Correcta:	15%
Respuesta en 5:	22%
Respuesta Aproximada:	4%
Fallos:	10.5%
No Respuesta:	63.5%

En un principio, esta visión podría resultar pesimista, si uno piensa que entre fallos y preguntas no respondidas suman el 74% de las preguntas, no obstante –sin ánimo de manejar los porcentajes como un político–, el elevado porcentaje de abstenciones es justificable:

- El sistema sólo puede contestar en función de los datos que tiene, es decir, si la solución a la pregunta no está en la fuente de información, el no responder no se considera fallo (más bien acierto). Sin embargo, no se dispone de los medios suficientes para mostrar cual es el porcentaje de preguntas del que no se tienen datos.

- Q^uA-C no es capaz de hacer inferencias, sólo mostrar frases de Wikipedia que se asemejen, de alguna manera, a las preguntas introducidas, por lo que preguntas como la 130 o 153 (mostradas en la **Figura 18**) no pueden ser contestadas.

- El sistema se diseñó para contestar oraciones interrogativas, lo que implica que preguntas como la 143 ("*Nombre los tres Beatles que siguen vivos*") no se interpretarán correctamente y, por lo tanto, tampoco se contestarán.

Tras estas consideraciones, los resultados del sistema no pueden considerarse malos, el nivel de aciertos no es muy elevado, 22%, pero es más de dos veces superior al de fallos, lo que no puede decirse de muchas personas.

6.4.2. Q^uA-C con expansión por sinónimos.

Una vez introducido el módulo de expansión por sinónimos en el sistema, se volvieron a realizar las doscientas preguntas, sin observarse ningún cambio relevante en el tiempo de procesado de la pregunta, entre 2 y 3 segundos, sin embargo, sí se observan diferencias importantes en los porcentajes:

Tabla 6. *Resultados de* Q^uA-C *con expansión por sinónimos.*

Respuesta Correcta:	16.5%
Respuesta en 5:	26.5%
Respuesta Aproximada:	5%
Fallos:	13.5%
No Respuesta:	55%

Como se puede comprobar, al incluir sinónimos en la búsqueda el sistema es capaz de responder un 8.5% más de preguntas; aunque no todo este porcentaje se contesta correctamente, el mayor aumento se produce en la categoría de aciertos, alrededor de un 4.5%.

En la **Figura 23** se muestra una de las preguntas (146: "*¿Cuándo tuvo lugar el referéndum para la adhesión de Noruega a la Unión Europea?*") que antes no tenían respuesta y que ahora, tras la inclusión del módulo con sinónimos, es contestada correctamente.

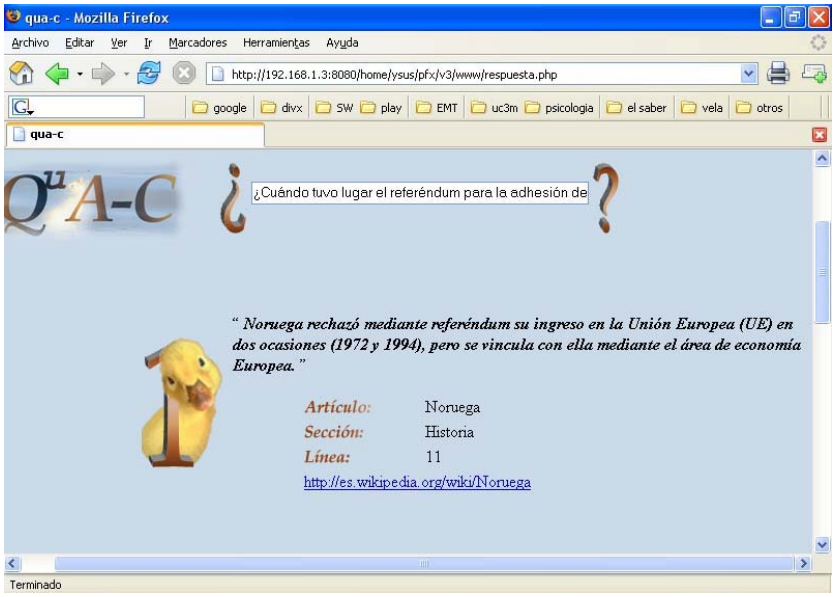

Figura 23. *Ejemplo de pregunta contestada con sinónimos (146).*

Este resultado hace valorar éste bloque positivamente, aunque el porcentaje de fallos también haya aumentado, lo hace en menor medida que el de aciertos; además, módulos como el de análisis del tipo de pregunta, que no actúan sobre el número de respuestas devueltas, se verán beneficiados por la adición de este bloque.

6.4.3. *Qu A-C* con análisis del tipo de pregunta.

Tras la unión del módulo de análisis del tipo de pregunta al sistema anterior, en el que se utilizaba la expansión por sinónimos, se evalúan de nuevo las capacidades del sistema, con el mismo procedimiento usado en los dos casos anteriores.

Con este módulo el número de preguntas respondidas se mantiene constante, debido a que su actuación no influye en el número de respuestas ofrecidas por el motor de búsqueda, sino en su ordenación; esta reordenación se hace notar, tanto en el aumento del número de "Respuestas correctas" (con un incremento del 2%) como en la disminución de los "Fallos" (en un 3%). La lista completa de los resultados obtenidos puede verse en la siguiente tabla.

Tabla 7. *Resultados de* ***Q^{u}A-C*** *con análisis del tipo de pregunta.*

Respuesta Correcta:	18.5%
Respuesta en 5:	27.5%
Respuesta Aproximada:	7%
Fallos:	10.5%
No Respuesta:	55%

A la vista de estos resultados, la inclusión de este bloque influye positivamente en los porcentajes de acierto, sin embargo, cabe hacer una importante distinción; puesto que, al igual que en los tiempos de respuesta, los resultados obtenidos en función de la categoría a la que pertenezca la respuesta son muy distintos:

- En el caso de que la respuesta sea considerada de tipo número, los resultados mejoran notablemente (el incremento del 2% en "Respuestas correctas" se debe casi exclusivamente a esta categoría). Este cambio puede observarse, por ejemplo, en las respuestas a la pregunta "*¿Cuándo se firmó el Tratado de Maastricht?*" (13) con módulo de análisis del tipo de pregunta, mostradas en la **Figura 24** y sin este módulo, mostradas en la **Figura 25**.

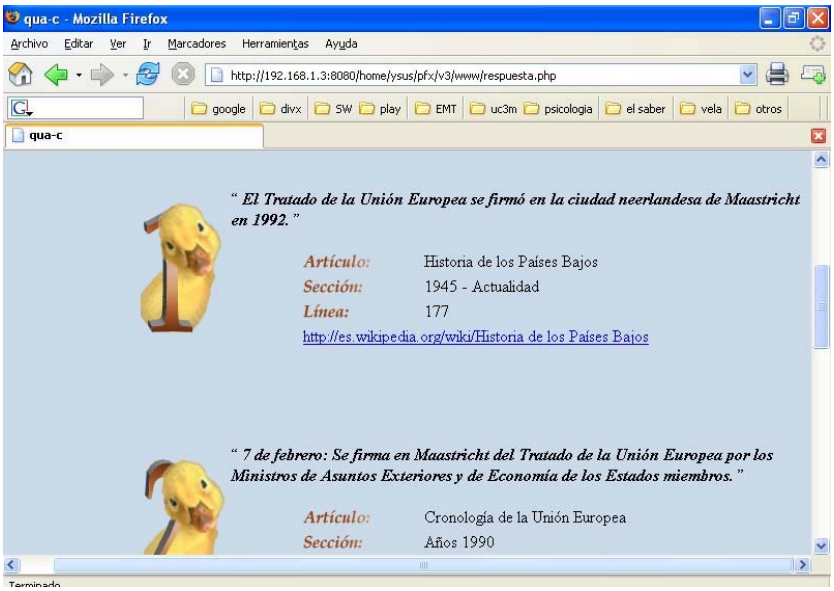

Figura 24. *Ejemplo de pregunta de tipo número con análisis del tipo (13).*

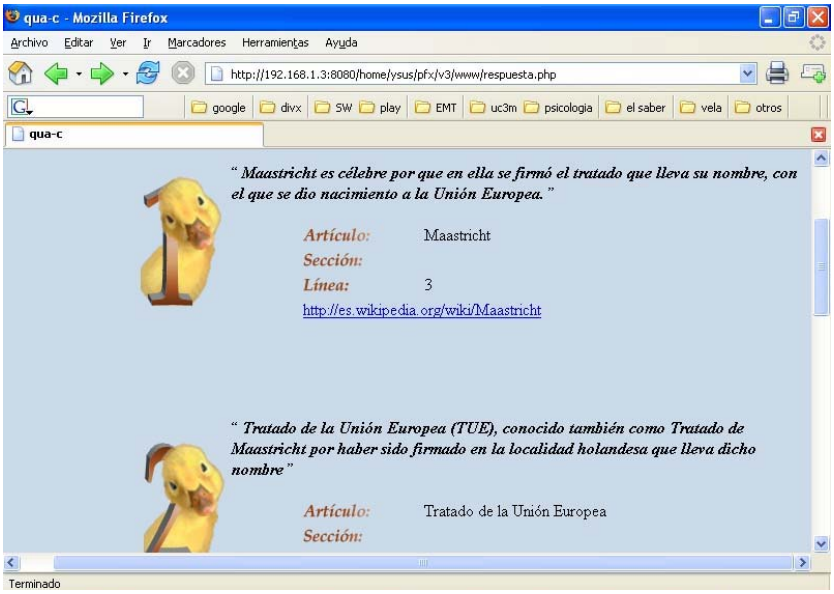

Figura 25. *Pregunta 13 sin análisis del tipo.*

- Sin embargo, en la categoría de nombre propio, los resultados obtenidos son muy similares a los que se obtenían sin la inclusión de este bloque. Esta diferencia se debe a que en la mayoría de las frases existe una palabra que puede clasificarse como nombre propio.

Otra característica reseñable es el aumento del tiempo de respuesta, en función del tipo de pregunta y del número de respuestas analizadas.

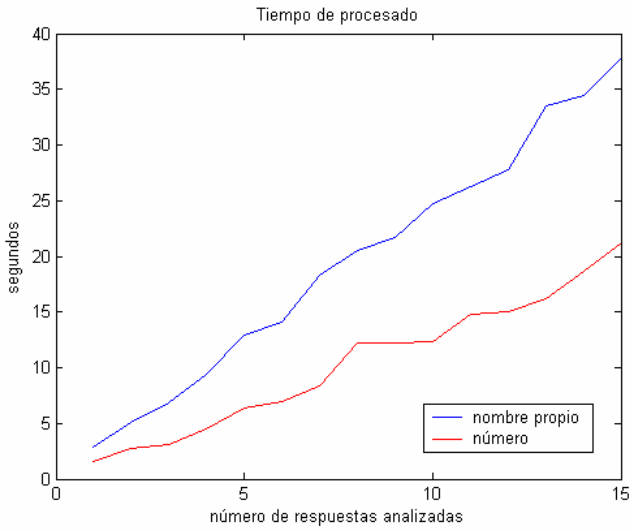

Figura 26. *Tiempo de procesado en función de la categoría de la respuesta.*

Como puede verse en la **Figura 26**, el aumento del tiempo de respuesta es función de dos factores:

- El tipo de pregunta, debido a que los procesos a los que se somete a las respuestas son distintos en función de la categoría.
- El número de respuestas procesadas, ya que el sistema irá analizando los fragmentos devueltos hasta encontrar 5 que se correspondan con el tipo de pregunta al que pertenecen.

Las ventajas que ofrece este bloque, por lo tanto, son relativas, y dependen de las categorías que se escojan. En este caso, se han obtenido muy buenos resultados al incluir una de las categorías y, sin embargo, no se han logrado progresos con la otra.

6.4.4. Comentarios finales.

Después de las pruebas realizadas a las distintas versiones del sistema, se ha demostrado que los dos bloques de expansión que se han añadido al sistema ofrecen mejoras visibles a los resultados del sistema básico. Sin embargo, habría que valorar de una manera más positiva el bloque de expansión por sinónimos, ya que la adición de este bloque, sin aumentar el tiempo de procesado, produce una disminución muy considerable del número de preguntas no contestadas, lo que, por una parte, aumenta el número de respuestas correctas; y por otra, incrementa el número de preguntas que otros bloques pueden tratar.

Capítulo 7.

Conclusiones y Trabajos Futuros

"[…] había una nota en la que, con letra de Elena, estaba escrita la frase: ‹‹Ningún sueño acaba como empieza››.*"*

Al despertar. **Daniel Rojo Pérez**

7.1. Conclusiones.

Se ha realizado una exhaustiva investigación sobre el estado del arte de los sistemas de QA, que ha permitido conocer los métodos y estrategias más novedosos que se utilizan actualmente en la construcción de estos sistemas.

A partir de ella, se ha desarrollado un sistema completo de QA que ha obtenido un nivel de aciertos más que aceptable, del 22%, y un tiempo de respuesta reducido, entre 2 y 3 segundos. Sin embargo, lo más destacable, aunque no positivo, de este sistema es el elevado número de preguntas no contestadas, el 63.5%; no obstante, este alto porcentaje puede ser debido, principalmente, a dos causas: el sistema no es omnisciente, está basado en una enciclopedia finita (Wikipedia), y sólo contestará cuando encuentre datos relacionados con la pregunta en los contenidos de ésta; y el sistema es incapaz de inferir, por lo que tampoco será capaz de devolver las respuestas que necesiten ser deducidas.

Mediante el uso de módulos de expansión se han mejorado las capacidades del sistema de manera visible, aunque no siempre en la medida que se esperaba.

La utilización del módulo de expansión por sinónimos, basado en el diccionario OpenThesaurus-es, ha conseguido reducir de manera importante (un 8.5%) el número de preguntas no contestadas, sin producir un aumento en el tiempo de procesado.

El segundo de los módulos estudiados, el de análisis de tipo de pregunta, aunque no muy extenso y sólo con tratamientos específicos para dos categorías de respuesta (número y nombre propio), ha provocado un importante aumento del tiempo de respuesta del sistema, de entre 5 y 10 segundos; no obstante, también ha ofrecido buenos resultados, aumentando en un 2% el número de respuestas correctas. Sin embargo, la aportación de cada categoría a este resultado es muy distinta, siendo prácticamente nula en el caso de nombre propio, debido a la existencia de este tipo de categoría morfológica en la práctica totalidad de los fragmentos devueltos por el motor de búsqueda.

7.2. Trabajos futuros.

Durante la implementación del sistema se han descubierto puntos dentro de su esquema a partir de los cuales se podrían originar líneas de desarrollo.

En la etapa de filtrado de los datos de Wikipedia aparecen algunas de ellas: la primera podría ser el tratamiento de los elementos de tipo tabla, en la versión actual, aunque contienen información, no son procesados, debido a la dificultad que conlleva tanto su interpretación (contienen mucha información de formato) como el tratado posterior de su contenido, ya que para poder indexar este tipo de datos correctamente se deben guardar referencias, no solo de los contenidos de la celda, si no de su significado; también se podría continuar mejorando esta etapa reduciendo el tiempo de procesado que, aunque ha ido descendiendo durante la evolución del sistema, aún es bastante elevado (unas 15 horas).

El diseño modular del sistema ofrece la capacidad de cambiar, fácilmente, algunos de sus componentes; esta posibilidad abre las puertas a investigaciones que evalúen las capacidades del sistema en función de los componentes que se utilicen, por ejemplo, en función del motor de búsqueda.

El módulo de análisis del tipo de pregunta utilizado en el sistema es todavía sólo un esbozo, en él se deberían crear un mayor número de categorías, diseñar los correspondientes tratamientos y evaluar su funcionalidad.

También sería conveniente añadir un módulo de realimentación al sistema, que le hiciera capaz de realizar distintos tratamientos en función del número de respuestas devueltas, para que, por ejemplo, en caso de no devolver ninguna respuesta, el sistema vuelva a realizar la búsqueda utilizando un mayor número de palabras "vacías".

Y la inclusión de un módulo de extracción de la respuesta más exacto, que devolviera respuestas concretas en vez de frases, teniendo en cuenta los buenos resultados obtenidos, haría posible la participación de este sistema en las convenciones internacionales como CLEF.

Apéndice A

Preguntas QA@CLEF 2006

"¿No vas a decirme que piensas quedarte sentado y dejar que […] te convenza de que eres un conejo?"

Alguien voló sobre el nido del cuco. **Ken Kesey**

0001 *¿Qué es el Atlantis?*
0002 *¿Qué es el Hubble?*
0003 *¿Qué es Nike Zeus?*
0004 *¿Qué es Linux?*
0005 *¿Quién es Iosif Kobzon?*
0006 *¿A qué país invadió Irak en 1990?*
0007 *¿Cuántos países forman la OTAN actualmente?*
0008 *¿Qué organización dirige Yaser Arafat?*
0009 *¿A qué organización desea pertenecer Taiwán?*
0010 *Nombre una película en la que haya participado Kirk Douglas en el periodo de 1946 a 1960.*
0011 *Dé el nombre de alguien que haya ganado el Premio Nobel de Literatura entre 1945 y 1990.*
0012 *¿Cuándo fue la coronación oficial de Isabel II?*
0013 *¿Cuándo se firmó el Tratado de Maastricht?*
0014 *¿Cuándo murió Stalin?*
0015 *¿Quién era el protagonista de la película "Siete años en el Tíbet"?*
0016 *¿Quién escribió la novela fantástica titulada "El señor de los anillos"?*
0017 *¿Cómo se llama la primera mujer que escaló el Everest sin oxígeno?*
0018 *¿Quién era conocido como el "Canciller de Hierro"?*
0019 *¿En qué año ocurrió la catástrofe de Chernobyl?*
0020 *¿En qué año nació Helmut Kohl?*
0021 *¿En qué año fue asesinado Martin Luther King?*
0022 *¿En qué país nació el Papa Juan Pablo II?*
0023 *¿A qué partido pertenece el primer ministro británico Tony Blair?*
0024 *¿Qué altura tiene el Kanchenjunga?*
0025 *¿Cómo se le llama también al Síndrome de Down?*
0026 *¿Dónde se celebraron los Juegos Olímpicos de Invierno de 1994?*
0027 *¿Qué organización ecologista se fundó en 1971?*
0028 *¿Para quién fingió trabajar entre 1970 y 1975 el director técnico del club de fútbol Bremen Willi Lenke?*
0029 *¿Qué altura tiene la Torre Eiffel?*
0030 *¿En qué estado americano está el Parque Nacional de Everglades?*
0031 *¿Quién descubrió el cometa Shoemaker-Levy?*
0032 *¿Con qué planeta chocó el cometa Shoemaker-Levy?*
0033 *¿Qué es la quinua?*
0034 *¿Qué empresa se hizo cargo de Barings después de su quiebra en Febrero de 1995?*
0035 *¿Quién es Nick Leeson?*
0036 *¿Quién era el presidente de Francia durante las pruebas de armas nucleares en el Pacífico Sur?*
0037 *¿Qué es la lepra?*

0038 *¿De qué organización es Peter Anderson el consejero en materia de alcohol?*
0039 *¿A qué partido político pertenecía Willy Brandt?*
0040 *¿Cuál es la palabra alemana más larga?*
0041 *¿Cómo se llama la moneda de Letonia?*
0042 *¿Qué es un GI Joe?*
0043 *¿Cuál es la principal religión de Timor Oriental?*
0044 *¿Cuántas casas se esperaban construir bajo la iniciativa Stirling entre 1993 y 1998?*
0045 *¿En qué ciudad dio positivo por estanozolol el corredor Ben Johnson durante los Juegos Olímpicos?*
0046 *¿Quién fue Alexander Graham Bell?*
0047 *¿Quién es Danuta Walesa?*
0048 *¿Quién es Vigdis Finnbogadottir?*
0049 *¿Quién es Marc Forné?*
0050 *¿Quién es Neil Armstrong?*
0051 *¿Quién es Fernando Masone?*
0052 *¿Quién es Javier Clemente?*
0053 *¿Qué es Lufthansa?*
0054 *¿Qué es Médicos Mundi?*
0055 *¿Qué es Airbus?*
0056 *¿Qué es el BIRF?*
0057 *¿Qué es Christie's?*
0058 *¿Qué es el CERN?*
0059 *¿Qué es Deep Blue?*
0060 *¿Qué es el tóner?*
0061 *¿Qué es Eurovisión?*
0062 *¿Qué es el Big Bang?*
0063 *¿Quiénes fueron los cosacos?*
0064 *¿Qué es el ECU verde?*
0065 *¿Qué es el dracma?*
0066 *¿Qué premiado por el Instituto Goethe no recogió el premio?*
0067 *¿Quién preside la RAI?*
0068 *¿Quién ganó la Batalla de El Alamein?*
0069 *¿Quién es el director de operaciones de la NASA?*
0070 *¿Quién es el presidente de Letonia?*
0071 *¿A quién se conoce como la "Dama de Hierro" de Hong Kong?*
0072 *¿Quién es el secretario general de la Interpol?*
0073 *¿Qué médico acompañó a Miguel Induráin en su desplazamiento a Colorado?*
0074 *¿Quién descubrió el ácido acetilsalicílico?*
0075 *¿En qué año se celebró el mundial de fútbol de Estados Unidos?*
0076 *¿En qué año se hundió el Titanic?*
0077 *¿Entre qué años tuvo lugar la Segunda Guerra Mundial?*
0078 *¿En qué fecha Estados Unidos invadió Haití?*
0079 *¿Qué día firmaron Jordania e Israel un acuerdo de paz?*
0080 *¿En qué año murió Bernard Montgomery?*
0081 *¿Cuándo se lanzó el telescopio Hubble?*
0082 *¿En qué año fue la revolución rusa?*
0083 *¿En qué año fue la retirada de Dunquerque?*
0084 *¿En qué año ganó Einstein el Premio Nobel de Física?*
0085 *¿En qué ciudad está el Centro Espacial Johnson?*
0086 *¿En qué ciudad está el parque acuático Sea World?*
0087 *¿En qué ciudad está el teatro La Fenice?*
0088 *¿En qué calle vive el primer ministro británico?*
0089 *¿Qué ciudad andaluza deseaba celebrar los Juegos Olímpicos de 2004?*
0090 *¿En qué país está el aeropuerto de Nagoya?*
0091 *¿En qué ciudad se celebró la 63 edición de los Oscar?*
0092 *¿Dónde está la sede de la Interpol?*
0093 *¿Dónde trabajaron juntos Braque y Picasso?*
0094 *¿Qué organismo realizó un llamamiento a la "Tregua Olímpica"?*
0095 *¿De qué organización fue director general Jacques Diuf?*
0096 *¿De qué organismo es director gerente Michel Camdessus?*
0097 *¿De qué organización es César Gaviria secretario general?*
0098 *¿De qué organización fue secretario general en funciones Sergio Balanzino?*
0099 *¿Cómo se llama la compañía alemana que comercializa los potitos de Hero?*
0100 *¿Qué organización mantiene un embargo sobre Irak?*
0101 *¿De qué estudios cinematográficos fue director artístico Cedric Gibbons?*
0102 *¿A qué grupo pertenece AVIACO?*
0103 *¿De qué sociedad ha sido miembro del Comité Ejecutivo Martín Bustamante?*
0104 *¿Cuántos habitantes tiene Longyearbyen?*
0105 *¿Cuántas piezas tiene el Tesoro del Carambolo?*

0106 *¿Cuántos Oscar ganó La guerra de las Galaxias?*
0107 *¿Cuántos soldados tiene España?*
0108 *¿Cuántos campeonatos de Formula Uno ganó Fangio?*
0109 *¿Cuántas categorías tienen los premios Grammy?*
0110 *¿Cuántos premios Grammy ganó El rey león?*
0111 *¿Cuánto dinero gana anualmente el narcotráfico?*
0112 *¿Cuál es el presupuesto de la Interpol?*
0113 *¿Cómo se llamó al primer submarino nuclear?*
0114 *¿En qué cárcel estuvo Mario Conde?*
0115 *¿Cuál es el componente principal de la Aspirina?*
0116 *¿Cómo se llama el grabado más grande de Picasso?*
0117 *¿Qué escudería preside Luca Cordero Di Montezemolo?*
0118 *¿Qué robaba el oso Yogi?*
0119 *¿Cuál es la especie más emblemática de Doñana?*
0120 *¿Cómo se llama la bicicleta con la que Miguel Induráin batió el record de la hora?*
0121 *¿Quién fue el primer presidente de Estados Unidos que visitó China?*
0122 *¿Quién fue el presidente de Perú entre 1985 y 1990?*
0123 *¿Quién ganó el Tour Francia de 1988?*
0124 *¿Qué zar ruso murió en 1584?*
0125 *¿En qué ciudad se celebró el partido inaugural del mundial de fútbol de Estados Unidos?*
0126 *¿De qué puerto partió el portaviones Eisenhower cuando se dirigió a Haití?*
0127 *¿Qué país presidió Roosevelt durante la Segunda Guerra Mundial?*
0128 *Nombre un país que se independizara en 1918.*
0129 *¿Qué organismo presidió Simón Peres después de morir Isaac Rabin?*
0130 *¿Qué organismo presidía Primo Nebiolo durante los Campeonatos del Mundo de atletismo de Gotemburgo?*
0131 *¿De qué cuerpo fue director Luis Roldán de 1986 a 1993?*
0132 *¿En qué organización entró Grecia en 1952?*
0133 *¿Cuántos años tenía Umberto Bossi cuando dejó de ser secretario general de la Liga Norte?*
0134 *¿Cuántos kilómetros se recorrieron en el tour de 1926?*
0135 *¿Cuántos países visitó Nixon entre 1953 y 1959?*
0136 *¿Cuántos habitantes tenía Hong Kong en 1993?*
0137 *¿Qué cargo ocupaba Francois Mitterrand cuando ingresó en el hospital de Cochin?*
0138 *¿Cómo se llama la colección de pinturas que hizo Goya entre 1819 y 1823?*
0139 *¿Qué guerra tuvo lugar entre los años 1939 y 1945?*
0140 *¿En qué deporte venció Europa a América en 1987?*
0141 *¿Quiénes participaron en la Conferencia de Yalta?*
0142 *¿Qué países forman parte del Tratado de Libre Comercio de América del Norte?*
0143 *Nombre los tres Beatles que siguen vivos.*
0144 *Nombre tres Estados bálticos*
0145 *¿Cuáles son las tres repúblicas eslavas?*
0146 *¿Cuándo tuvo lugar el referéndum para la adhesión de Noruega a la Unión Europea?*
0147 *¿Cuándo se derribó el muro de Berlín?*
0148 *¿Qué es el LZ 129 Hindenburg?*
0149 *¿Cuándo ganó Tom Twyker el Premio Nobel de la Paz?*
0150 *¿Quién fue el primer ministro de Inglaterra antes de John Major?*
0151 *¿Cuándo fue la reunión del G7 en Halifax?*
0152 *¿Qué es la RKA?*
0153 *¿Cuántas veces ha ganado Zinedine Zidane el US Open?*
0154 *¿Quién es el astronauta que ha estado más tiempo en el espacio?*
0155 *¿Cuáles son los siete países más industrializados del mundo?*
0156 *¿Cuál es la profesión de Gianni Versace?*
0157 *¿Qué famoso evento francés se celebra el 11 de Noviembre?*
0158 *¿Cuántos telespectadores siguieron la final del mundial de fútbol de 1994?*
0159 *¿A cuántos periodistas hirió Maradona con un rifle de aire comprimido?*
0160 *¿Qué película ganó el Oso de Oro de 1988?*
0161 *¿Quién es Fernando Henrique Cardoso?*
0162 *¿Quién era el presidente del Deutsche Bank durante la quiebra del especulador inmobiliario Juergen Schneider?*
0163 *¿Qué multinacional francesa cambió su nombre por el de Grupo Danone?*
0164 *¿Quién es Rolf Ekeus?*
0165 *¿Qué países forman el Consejo de Cooperación del Golfo?*
0166 *¿Quién era George Starckmann?*
0167 *¿A cuánto asciende la multa que se le impuso a Italia por superar la cuota de producción de leche?*
0168 *¿Qué es la Asociación por la Paz?*
0169 *¿Cuántas nominaciones a los Oscar obtuvo "En el nombre del Padre"?*
0170 *¿Cuántas separaciones hubo en Noruega en 1992?*
0171 *¿Cuándo se celebró en Irlanda el referéndum sobre el divorcio?*

0172 *¿Quién fue el personaje favorito del museo de cera de Londres en 1995?*
0173 *¿Quiénes son los actores de "El color de la noche"?*
0174 *¿Cuándo se celebró en 1994 la 51 edición del Festival Internacional de Cine de Venecia?*
0175 *¿Cuándo se independizó Surinam?*
0176 *¿Cuál es el apodo de Eddy Merckx?*
0177 *¿Dónde está enterrado James Ensor?*
0178 *¿Dónde está el Hermitage?*
0179 *¿Dónde está situado Ystad?*
0180 *¿Quién creó el sistema operativo OS/2?*
0181 *¿Qué es el CBGB?*
0182 *¿Qué es el CD-i?*
0183 *¿En qué año murió Glenn Gould?*
0184 *¿Quién es Jan Tinbergen?*
0185 *Nombre luchadores de sumo.*
0186 *¿En qué ciudad de Zelanda pasaba varias semanas al año Jan Toorop entre 1903 y 1924?*
0187 *¿Qué es un samovar?*
0188 *¿Qué es la "Bundesgrenzschutz"?*
0189 *¿Qué es Roque Santeiro?*
0190 *¿Qué es la cachupa?*
0191 *¿Qué fue anexionado después de la Guerra de los Seis Días?*
0192 *¿A qué estado pertenece Porto Alegre?*
0193 *¿Qué país invadió Gran Hanish?*
0194 *¿A cuánto ascendió la multa a John Fashanu?*
0195 *¿Cuál es el record del mundo de salto de altura?*
0196 *¿Cómo se llama la compañía de ferrocarriles francesa?*
0197 *¿Cuál fue el resultado del partido Italia-Nigeria de la Copa del Mundo de 1994?*
0198 *¿Cuál es la nacionalidad de Geoffrey Oryema?*
0199 *¿Qué enseña Vital do Rego?*
0200 *¿Desde cuándo Portugal es una república?*

Bibliografía*

"Se pasan el tiempo mirando fijamente su contenido. Llaman a eso "leer"[…] Los de Artículos de Escritorio se los reservan para ellos -explicó Dorcas-. Dicen que los libros le hinchan a uno el cerebro, a menos que sepa leerlos como es debido.

La Nave. **Terry Pratchett**

[1] Atserias, J., B. Casas, E. Comelles, M.González, L. Padró y M. Padró.
Freeling 1.3: Syntactic and semantic services in an open-source NLP library. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06). Genoa, Italy. 2006.*
http://www.lsi.upc.es/~bcasas/publications/
más referencias de Freeling:
http://garraf.epsevg.upc.es/freeling/index.php?option=com_content&task=view&id=20&Itemid=49

[2] Atserias, Jordi, Carmona, Josep, Castellón, Irene, Cervell, Sergi, Civil, Montserrat, Màrquez, Lluís, Martí, Mª Antònia, Padró, Lluís, Placer, Roberto, Rodríguez, Horacio, Taulé, Mariona y Turmo, Jordi.
Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text. *Proceedings of the 1st International Conference on Language Resources and Evaluation, LREC, pg. 1267--1274.Granada, Spain. May, 1998.*
http://www.lsi.upc.es/~nlp/papers/1998/lrec98-a.al.ps.gz

[3] Atserias, Jordi, Comelles, Eli, Mayor Aingeru.
Txala: un analizador libre de dependencias para el castellano. *Procesamiento del Lenguaje Natural, (35):455–456, Septiembre 2005.*
http://www.lsi.upc.edu/~comelles/Publicacions/sepln05demo.pdf

[4] Atserias, Jordi y Rodríguez Horacio.
TACAT: Tagged corpus analyzer tool. *Technical report lsi-98-2-t, Departament de LSI. Universitat Politècnica de Catalunya.*
http://citeseer.ist.psu.edu/262646.html

[5] Bikel, Daniel M., Schwartz, Richard y Weischedel, Ralph M.
An Algorithm that Learns What's in a Name. *1999.*
http://www.cis.upenn.edu/~dbikel/papers/algthatlearns.doc.pdf

* Todos los enlaces que se muestran junto a cada una de las referencias bibliográficas han sido comprobados el 2 de julio de 2006.

[6] Brants, Thorsten.
Tnt – a statistical part–of –speech tagger. *In Proceedings of the 6th Conference on Applied Natural Language Processing. ANLP. ACL. 2000.*
http://citeseer.ist.psu.edu/brants00tnt.html

[7] Burger, John y Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, Sanda Harabagiu, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, Dan Moldovan, Bill Ogden, John Prager, Ellen Riloff, Amit Singhal, Rohini Shrihari, Tomek Strzalkowski, Ellen Voorhees y Ralph Weishedel**.**
Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q & A)
http://www-nlpir.nist.gov/projects/duc/papers/qa.Roadmap-paper_v2.doc

[8] Carlos Arias Fernandez.
Aprovisionamiento Inicial en Pasarelas de Servicios*. Proyecto Fin de Carrera, diciembre de 2005*

[9] Daelemans, Walter, Zavrel Jakub, Berk, Peter y Gillis, Steven.
Learning Parse and Translation Decisions from Examples with Rich Context. *E. Ejerhed and I. Dagan (eds.) Copenhagen, Denmark, 14-27, 1996.*
http://acl.ldc.upenn.edu/W/W96/W96-0102.pdf

[10] Gómez Torrego, Leonardo.
Gramática didáctica del Español. *De la octava edición (páginas 97, 126 y 208). Ediciones SM, Madrid, enero de 2002.*

[11] Goñi-Menoyo, José M., González, José C., Martínez-Fernández, José L., Villena-Román, Julio, García-Serrano, Ana M., Martínez-Fernández, Paloma, de Pablo-Sánchez, César y Alonso Sánchez, Javier.
MIRACLE's hybrid approach to bilingual and monolingual Information Retrieval. *Clef 2004.*
http://www.clef-campaign.org/2004/working_notes/WorkingNotes2004/19.pdf

[12] Hermjakob, U. y R.J.Mooney.
Learning Parse and Translation Decisions from Examples with Rich Context. *In 35th Proceedings of the Conference of the Association for Computational Linguistics (ACL), 482-489.*
http://citeseer.ist.psu.edu/cache/papers/cs/27816/http:zSzzSzacl.ldc.upenn.eduzSzPzSzP97zSzP97-1062.pdf/ulf97learning.pdf

[13] Llopis, Fernando y Vicedo, José Luis.
IR-n, a passage retrieval system from University of Alicante. *Clef 2001.*
http://www.ercim.org/publication/ws-proceedings/CLEF2/llopis.pdf

[14] Miller, George A. y Christine Fellbaum, Randee Tengi, Susanne Wolff, Pamela Wakefield, Helen Langone y Benjamin Haskell.
Wordnet
http://wordnet.princeton.edu/

[15] P. Daniel Borches Juzgado.
Monitorización Remota de Pacientes con Problemas Cardiacos: Transmisión y Visualización Remota de un ECG a través de Bluetooth y GPRS/UMTS. *Proyecto Fin de Carrera, septiembre de 2005.*

[16] Rabinwitz, Josh.
How to Index Anything. *LinuxJournal.com. Julio 2003.*
http://joshr.com/src/docs/HowToIndexAnything.pdf

[17] Shimmin, Tim.
MG Overview Notes
http://www.mds.rmit.edu.au/mg/prog_notes/sys_overview.html

[18] Tiedemann, Jörg.
A comparison of off-the-shelf IR engines for question answering. *Alfa-Informatica, University of Groningen, The Netherlands.*
http://www.let.rug.nl/~gosse/Imix/clin04_tiedemann.pdf

[19] University of Amsterdam
EuroWordNet
http://www.illc.uva.nl/EuroWordNed/

Glosario

"El primero de la estirpe está amarrado en un árbol
y al último se lo están comiendo las hormigas."

Cien años de soledad. **Gabriel García Márquez**

ASCII: "American Standard Code for Information lnterchange"

CLEF: "Cross-Language Evaluation Forum"

GDFL: "GNU Free Documentation Licence"

GNU: "GNU's Not UNIX"

HTML: "HyperText Markup Language"

IR: "Information Retrieval" (en español, RI)

LGPL: "Lesser General Public Licence"

NLP: "Natural Language Processing" (en español, PLN)

PHP: "PHP: Hypertext Preprocessor"

PLN: "Procesamiento del Lenguaje Natural" (en inglés, NLP)

QA: "Question Answering"

Q"AC: "Question Answering – Carlos III"

RI: "Recuperación de Información" (en inglés, IR)

TREC: "Text REtrieval Conference"

UTF8: "8-bit Unicode Transformation Format"

XML: "Extensible Mark-up Language"